Towards Interactive Guidance for Writing Training Utterances for Conversational Agents

David Piorkowski
djp@ibm.com
IBM Research
Yorktown Heights, NY, USA

Rachel Ostrand
rachel.ostrand@ibm.com
IBM Research
Yorktown Heights, NY, USA

Kristina Brimijoin
kbrimij@us.ibm.com
IBM Research
Yorktown Heights, NY, USA

Jessica He
jessicahe@ibm.com
IBM Research
Seattle, WA, USA

Erica Albert
erica.albert@ibm.com
IBM
Denver, CO, USA

Stephanie Houde
stephanie.houde@ibm.com
IBM Research
Cambridge, MA, USA

ABSTRACT

Improving conversational agents that are trained with supervised learning requires iteratively refining example intent training utterances based on chat log data. The difficulty of this process hinges on the quality of the initial example utterances used to train the intent before it was first deployed. Creating new intents from scratch, when conversation logs are not yet available, has many challenges. We interviewed experienced conversational agent intent trainers to better understand challenges they face when creating new intents, and their best practices for writing high quality training utterances. Using these findings and related literature, we developed an intent training tool that provided interactive guidance via either language feedback or sample utterances. Language feedback notified the user when training utterances could be linguistically improved, while sample utterances were crowdsourced and provided examples of end user language prior to deploying an intent. We compared these two types of guidance in a 187-participant between-subject study. We found that participants in the language feedback condition reported limited creativity and higher mental load and spent more time on the task, but were more thoughtful in crafting utterances that adhered to best practices. In contrast, sample utterance participants leveraged the samples to either quickly select examples or use them as a springboard to develop new utterance ideas. We report on differences in user experience in the strategies that participants took and preferences for or against the different types of guidance.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in collaborative and social computing**.

KEYWORDS

artificial intelligence; machine learning; AI model development; AI explainability

CUI '24, July 08–10, 2024, Luxembourg, Luxembourg

ACM ISBN 979-8-4007-0511-3/24/07
https://doi.org/10.1145/3640794.3665553

ACM Reference Format:
David Piorkowski, Rachel Ostrand, Kristina Brimijoin, Jessica He, Erica Albert, and Stephanie Houde. 2024. Towards Interactive Guidance for Writing Training Utterances for Conversational Agents. In *ACM Conversational User Interfaces 2024 (CUI '24), July 08–10, 2024, Luxembourg, Luxembourg*. ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3640794.3665553

1 INTRODUCTION

Conversational agents continue to be ubiquitous across many diverse domains. Traditional conversational agents are typically trained using supervised learning techniques and multiple *example utterances* (the text that resembles users' inputs) for each of their *intents* (the topics the agent should recognize). More recently, conversational agents trained on large language models (LLMs) are rising in popularity. However, such agents have disadvantages including higher costs to run, less determinism, and hallucinations - responses with incorrect or misleading information. Already, the consequences of relying on LLM-generated output are being reported. For example, Air Canada was recently required to honor a partial refund based on false information provided by their conversational agent [2] and New York City's "MyCity" agent was caught lying about laws and regulations [31]. As such, traditional conversational agents may be strongly preferable in certain domains, for their small data requirements and the finer control and explainability of agent output they offer, particularly in high-risk applications or regulated industries such as banking or insurance.

Instead of using LLMs for output like the examples above, a conversational agent could use an LLM for intent classification, and rely on more deterministic solutions for generating output. Classification approaches like zero-shot learning or few-shot learning have been shown to be highly accurate and low-effort ways to classify text in common domains [34]. Although LLMs are very promising, they too have limitations. Prior work has shown that classification quality can decrease with specific domains [30], a large number of classes [51], classes that are insufficiently descriptive [48], or data that has not been human-curated [23]. In such instances, fine tuning has emerged as a technique to augment the training data used by an LLM to specialize it to a particular task. However, fine-tuning approaches require a large number of well-labeled examples, similar to the training requirements of traditional conversational agents [32]. Furthermore, fine tuning is challenging and if done

improperly, it may further degrade classification quality [7]. Additionally, recent work suggests that considering factors such as model performance, timing, energy use, and cost may holistically favor traditional approaches over LLMs for certain tasks [36]. LLMs are clearly going to continue to improve and play a prominent role in the development of conversational agents moving forward. However, for highly specific domains (which would require additional data for LLM fine tuning), in cost- or energy-constrained contexts, or instances where oversight and explainability are important, traditional conversational agents may be more suitable.

Traditional conversational agents are not without challenges, however. The process of creating or selecting example utterances for training is known to be time-consuming and error prone, and often requires substantial iteration to reach an adequate level of performance [4, 45]. However, some well-established tools and improvement practices are available to address some of these issues [21]. For example, IBM's Watson Assistant offers best practice guidelines that suggest starting with 10-20 examples per intent and gathering example utterances from log data after deployment to further improve intent recognition [3].

Traditional conversational agent models will need to be retrained many times in order to better converge on the needs of end users [1, 18]. Post-deployment *usage logs* - lists of utterances that real end users have sent to the agent - are used in the retraining to improve the quality of the intent recognition [16, 38]. Tools like watsonx Assistant and Microsoft's Azure Bot address this challenge by offering analytics to identify problematic interactions and poorly trained intents.

However, the aforementioned tools and guidelines do little to aid conversational agent developers in creating a new conversational agent from scratch or adding a new intent to an existing one. A conversational agent developer needs to craft sufficient representative examples so that the machine learning algorithm that is driving the conversational agent has enough information to identify similar utterances provided by a broad spectrum of end users. This means the examples must be diverse enough to account for an extensive range of utterances [47, 52], and the developer needs to account for this variety in the examples they craft. They additionally must consider how new examples may interact with examples for other intents that the agent has already been trained on to avoid conflicts between similar examples. Furthermore, developers who are new to training conversational agents may not have learned the nuances of training within a particular conversational system for a particular domain, including the techniques that more experienced developers have identified for achieving the best performance for a particular use and context.

To address these challenges in the development of a new conversational agent, we hypothesized that it would be helpful for conversational agent developers who write training utterances for new intents to receive in-tool guidance. Based on the current work practices and needs of utterance writers, we designed a conversational agent intent training tool to test two types of guidance: one that provides language feedback and another that provides sample utterances. The focus of this paper is evaluating the impact of these two forms of guidance on novice utterance writers engaged in training a new intent. We investigate this guidance in a user study which compared participants' performance on the following three conditions:

- *Baseline*, which provided basic dos and don'ts for intent training as static documentation alongside the tool, mimicking existing guidance used in practice today.
- *Language Feedback*, which provided timely, interactive feedback on specific linguistic properties of utterances that impact intent training.
- *Sample Utterances*, which made a list of user-generated utterances available to utterance writers prior to the deployment of the conversational agent.

We conducted a user study to evaluate these forms of guidance for conversational agent intent training, seeking to answer the following research questions:

- RQ1. How does each form of guidance affect the quality of training utterances written by novices?
- RQ2. How does each form of guidance impact the user experience of novice utterance writers?

2 RELATED WORK

2.1 Supporting Conversational Agent Training

Although there is prior work which focuses on supporting various design aspects of conversational agents such as agents' personalities [28], their conversational behavior [40], or addressing specific problems or users [15, 19, 24, 39], there is limited research investigating how to support the *training* of conversational agents, especially when log data is not yet available. Prior work has found that practitioners who construct conversational agents tend to lack technical knowledge of how conversational systems work, including among those who are involved in the implementation of the natural language understanding in the system [43].

Approaches exist to support the creation of conversational agents outside of writing training utterances. For example, several techniques are available to identify candidate intents from existing conversational log data [14, 22, 25, 41]. Instead of helping to create utterances from scratch for a given intent, these techniques provide intents from a set of existing utterances and can serve as a starting point for further refinement. This approach can be extended to consider the larger conversational development process, including evaluation. For example, Williams and colleagues [49] proposed an interactive learning approach to building conversational agents that "combines model definition, labeling, model building, active learning, model evaluation, and feature engineering in a way that allows a domain expert ... to build classifiers". Like the automated approaches, log data is a prerequisite. Pérez-Soler and colleagues assessed and catalogued 14 well-known conversational agent development platforms, and explored end-to-end technical factors including sentiment analysis, NLP for phrase matching and text processing [33]. They found that none of the platforms, frameworks or libraries offered design patterns or quality metrics, just minimal support in the form of informal guidelines. Our work aims to address this gap by guiding the training of new intents, particularly when log data is not available.

Candello and colleagues studied the building and debugging practices of conversational agent knowledge workers [5]. Their research explained how knowledge workers identified intent-related problems by searching for problematic utterances from existing logs. The workers' task centered on first determining whether a problem was caused by intent collision (when multiple distinct intents have overlap in the topics they address) or the need for a new intent, and then editing, deleting or adding examples accordingly. Candello et al. found that tools to support knowledge workers' needs to perform these tasks ranged from incomplete to non-existent. The paper noted many challenges in the utterance-writing process, including participants having low visibility into the conversational agent's understanding of their utterances. To help address the challenges they identified, the authors developed a set of design implications for conversational platforms, including guides on language tone and features to track semantic and syntactic elements of the utterances which caused breakdowns or errors in the conversational agent.

The semantic and syntactic failures identified in that work was the starting point of the interface design and guidance tool that we developed in the current study, to include lexical and syntactic variety information. Addressing these issues during the initial training potentially improves the quality of the agent and reduces the amount of iteration needed in spoken language systems [8]. We seek to expand on this work by providing human-centered means for improving linguistic diversity during intent training for conversational agents.

2.2 Evaluating Conversational Agents

Knowing what makes a good or effective conversational agent is challenging, but is important to define when engaged in training. Traditional algorithmic evaluations of conversational agents include accuracy, precision, F1 score, precision-recall (PR) and receiver operating characteristic (ROC) curves. But human perception of the conversational agent is much more complex than those numeric metrics can encapsulate. Several studies explore this topic. A literature review by Radziwill and Benton [35] found that the International Organization for Standardization's (ISO) dimensions of effectiveness, efficiency, and satisfaction accurately apply to conversational agent quality. For effectiveness, papers cited in the review suggest combining subjective reviews with algorithmic evaluations along with iterative, comparative evaluations of different conversational agent versions [13]. Other authors suggest evaluating linguistic quality of agent responses [9, 46]. To address efficiency, one paper suggested performance test scripts to determine if the agent meets basic linguistic requirements while keeping confusions to a minimum [46]. Another paper suggested that for the satisfaction dimension, various qualitative traits of conversational agents should be assessed, including enjoyment, reduction in frustration, and ability to comment or otherwise provide feedback (among other characteristics), keeping in mind differing user goals across chat systems [27]. One commonality across these approaches is that they occur *after* the agent has been trained, which is arguably late in the development process. Our work is motivated by and builds on these approaches by providing feedback on quality *earlier* in the agent development process.

3 EXPLORATORY INTERVIEWS

As we encountered little previous research on the work practices of conversational agent developers, we first gathered expert insight via exploratory interviews with six professionals involved in the training process. Their roles included writing training utterances, developing and deploying conversational agents, testing the agents, and continuously improving the agents. The goal of these interviews was to inform and validate the design of our main experiment by understanding experts' work practices, pain points, what kinds of guidance would be valuable to them.

We provided participants with a simulated utterance writing environment via Mural [1] that incorporated assistance inspired by informal discussions with product managers of conversational agents, along with the pain points identified by Candello et al. [5]. The Mural included best practices, sample utterances, and the option to request a score that measured variety in the vocabulary and phrasing of their utterances. Variety and phrasing scores were on a scale of 'weak' to 'strong' and were chosen and simulated by the moderator. We asked participants to use the environment to write and/or select example training utterances for a new, hypothetical conversational agent intent while thinking out loud and discussing their current work practices.

We learned that training conversational agents is a highly collaborative, iterative task that requires expertise in how natural language understanding (NLU) training works. The majority of participants did not have machine learning backgrounds, so such expertise was often acquired by trial and error as the training process is, as one participant described, *"very blackbox."* Also contributing to the steep learning curve is that the best practices for writing training utterances vary greatly across conversational agents by content, domain and development platforms. Another participant said, *"If I have to train up a new person...it's about 3 to 6 months before they're actually adding value to the team."* The idea of interactive feedback customized to team practices was appealing to participants for reducing this learning curve.

We also learned that the initial set of training utterances usually does not accurately reflect end-user language, which is unpredictable due to differences in vocabulary and conversational styles across locales and individuals. This finding aligns with prior research [12, 44]. When asked about their current intent training practices, participants reported that conversational agent developers gather ideas for end user language through subject matter experts and customer support channels prior to deploying a new agent or skill, but they can only seek out real user utterances post-deployment via chat logs, resulting in a conversational agent being deployed and that does not capture the full range of user utterances. The sample utterances of the simulated environment mimicked such logs and was the feature most used by participants.

These exploratory interviews helped to validate our proposed guidance features and served as the basis for the two conditions in the main study. Language Feedback was motivated by a lack of such feedback in participants' existing tools. Sample Utterances was motivated by training utterances not reflecting user language. These are discussed in more detail in the following section.

[1] https://www.mural.co/

Table 1: The four intents used in the study.

Intent	Description
Lost or Stolen Card	Your card has been lost or stolen, and you need help with what to do.
Compromised Card	Someone else has used your card, and you need help with what to do.
Card Arrival	You want to know about your new card. This may include topics like status, arrival, delay, or others.
Order Physical Card	You don't have a physical card and want more information about ordering a new card. This may include topics like fees, delivery, process, or others.

4 MAIN STUDY: LANGUAGE FEEDBACK VS. SAMPLE UTTERANCES

Based on the findings from the exploratory interviews, we focused our main study on two kinds of writing-time guidance that could support novice conversational agent trainers: (1) interactive language feedback on utterances based on best practices for intent training, and (2) sample utterances made available prior to deployment.

4.1 Study Overview

4.2 Intent Training Tool

The main study was un-moderated, deployed as a web application, and consisted of four parts: the introduction, baseline task, experimental condition task, and post-experiment questionnaire.

In the introduction, participants were given a one-page primer that explained the experimental context for the conversational agent, what intent training was, and how example utterances are used to train the conversational agent to handle customer inquiries. Next, all participants completed the baseline condition by writing at least 10 utterances, using a baseline version of the tool containing neither feedback condition's features. In the experimental condition, participants were randomly assigned to either the *Language Feedback* condition or the *Sample Utterance* condition. Like the baseline condition, participants were required to provide a minimum of 10 utterances for the given intent, this time with the tool modified to offer participants utterance-writing guidance according to the condition they were assigned to (see Section 4.2 for tool details). After completing the two utterance-writing tasks, participants were directed to a web survey that asked them questions about their experiences with the baseline condition and their assigned experimental condition, the strategies they employed, and their preferences. The questions then assessed the perceived usability, pros and cons of each tool, as well as basic demographic and language history information.

The experiment tested a total of four intents that were randomly assigned across participants (see Table 1). Each participant received two of these intents, one for the baseline condition and one for the experimental condition. Assignment of experimental condition (*Language Feedback* vs. *Sample Utterance*), as well as which intent was displayed for each of the baseline and experimental conditions, was counterbalanced across participants, for a total of 24 unique combinations. The motivation for choosing four intents to counterbalance across participants was to reduce any possible effects of preference, specialized knowledge, or other forms of bias by participants. For example, having personal experience with one of the intents could lead participants to write more varied and a greater number of utterances for that intent, irrespective of the guidance condition. Due to participants who launched the tool but did not complete the study, the final set of included participants was not perfectly balanced across all combinations of intents and conditions. However, the number of participants in each group was relatively balanced, with 19–28 participants in each combination of experimental condition and intent presented for that condition.

4.3 Intent Selection

When designing the intents for use in the study, we wished to select a domain that would be broadly familiar to participants, as well as one that has a publicly available intent classification dataset on which to train the conversational agents. We used the *Banking77* dataset [6], a freely-available dataset which consists of 13,083 curated customer service queries, each labeled into one of 77 intents in the banking sector. We selected *Banking77* because it contains 77 fine-grained topics of banking-related customer service queries. The dataset is regularly used to evaluate models in natural language classification work [17, 20, 23] and its large dataset size, topic familiarity, and chat-oriented nature served the needs for our study. We chose four bank card-related intents to use in our study (see Table 1) that we expected would be commonly understood by our participant pool, with overlapping vocabulary between the different intents such that the classification task would not be extremely straightforward.

To evaluate the two kinds of guidance, we built an intent training tool (Figure 1) where participants could craft utterances for a given intent. The tool had the following interface features across all conditions of the experiment: (A) name and description of the intent, (B) the utterance editor for adding, editing, and deleting training utterances, and (C) a panel that described dos and don'ts for writing training utterances, along with the conversational agent training basics provided in the study introduction.

4.3.1 Baseline Condition. The tool in the *Baseline* condition had no support to aid participants in writing training utterances except the guidelines in the rightmost panel (Figure 1-C). Guidelines included varying the terms and phrasing used across utterances, avoiding filler words (e.g., greetings, chit chat, and words like "please" or "thanks"), only including one intent per utterance, and keeping utterances concise (i.e., under 120 characters). These guidelines were synthesized from best practices shared by conversational agent developers in the exploratory interviews.

4.3.2 Language Feedback Condition. In the *Language Feedback* study condition, the intent training tool included additional in-situ language-based feedback that appeared as the user added new utterances (Figure 2-Top). The types of feedback provided to participants

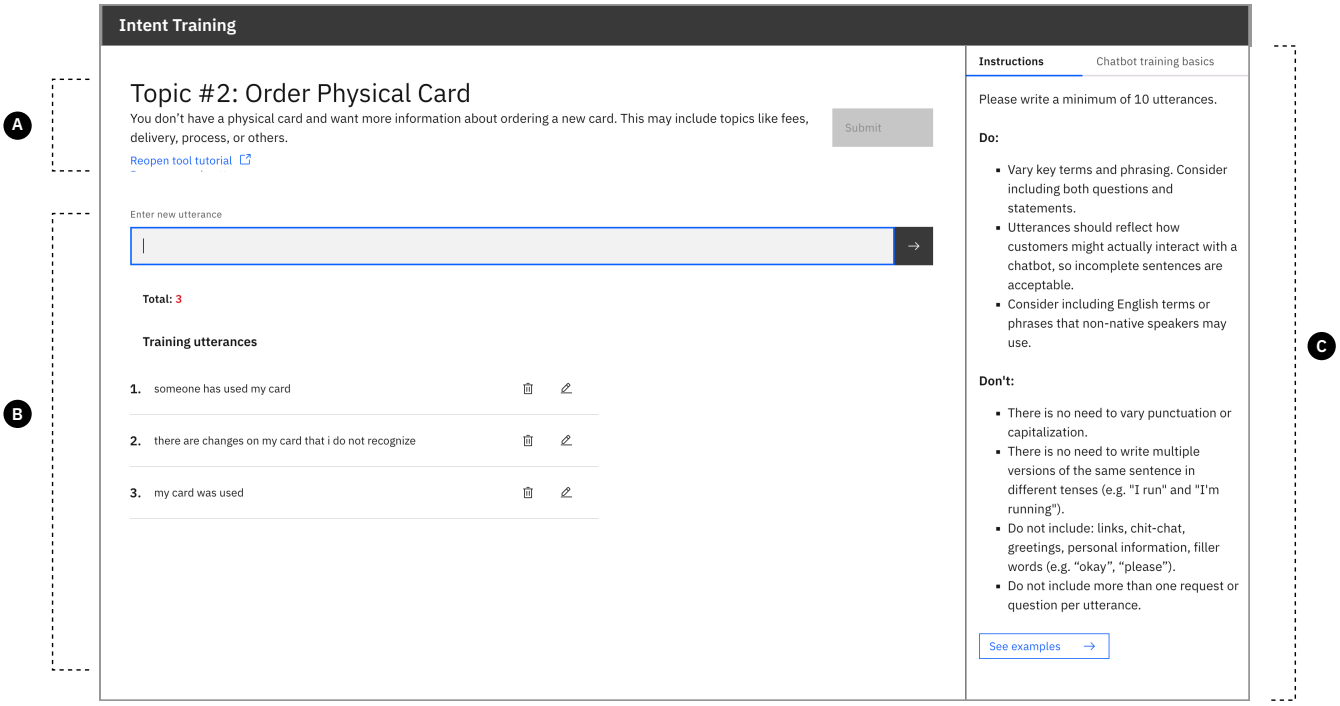

Figure 1: The intent training tool in the Baseline condition. The main features of the tool include the current intent and description (A), the utterance editor (B), and a help panel containing guidelines in the form of Do and Don't lists and a brief description of how conversational agent training works (C).

are shown in Table 2 and were also based on best practices shared by conversational agent developers in the exploratory interviews.

For Term Similarity, we first removed stop words and stemmed words using the Porter stemmer algorithm [2]. We then calculated the term frequency inverse document frequency (TF-IDF) vector for each utterance, using all of the participants' added utterances as the set of documents for the inverse frequency part of the calculation. Two utterances were flagged as similar when the cosine similarity of the pair's vectors was greater than 0.7. Phrase similarity was measured using Brill's Part of Speech tagger [3] and finding the largest common sub-sequence of tags between each pair of utterances. If the length of the largest common tag sub-sequence was greater than 60% of the number of words in the sentence, the utterances were flagged as similar.

The tool provided feedback notifications adjacent to training utterances that met a notification threshold. Each utterance's feedback updated every time the participant added a new utterance or edited an existing one. Hovering over a feedback notification highlighted the word(s) that contributed to the feedback. For example, hovering over the *similar terms* notification highlighted the similar terms across *all* utterances. In addition to per-utterance feedback, the tool also displayed overall measures of lexical variety (called *term variety* in the tool) and syntactic variety (called *phrase variety* in the tool) that also updated with each addition or edit to the utterances. Participants were shown a tutorial that described the different kinds of feedback prior to beginning the task with Language Feedback. They were not required to resolve feedback before proceeding to the next step of the study.

4.3.3 Sample Utterance Condition. In the *Sample Utterance* condition, the intent training tool additionally included a widget containing a searchable, multi-page list of approximately 100 sample utterances for the given intent as shown in Figure 2-Bottom. From this widget, participants could copy and paste an utterance into their set or look for inspiration to write their own utterances—the tool intentionally did not specify if or how the samples should be used. This condition explored the idea of making sample utterances available to utterance writers *prior* to deploying a conversational agent, in response to exploratory interview findings on the limitations of only getting user-generated utterances through chat logs *post*-deployment. Since we did not want to favor the intent classfier with examples from *Banking77's* training or test dataset, we turned to crowdsourcing. We gathered user-generated sample utterances through an internal crowdsourcing survey, in which business users were given a scenario based on the sample intent and asked, *"How would you ask the chatbot to help you?"* Survey participants provided 5-10 utterances for each of the four intents used in the main study. Twenty participants contributed a total of 433 sample utterances across the four intents (range: 105–112 utterances per intent across participants; 20–34 utterances written per participant). People who participated in this survey were not eligible for participation in the main study.

[2] https://www.nltk.org/_modules/nltk/stem/porter.html
[3] https://naturalnode.github.io/natural/brill_pos_tagger.html

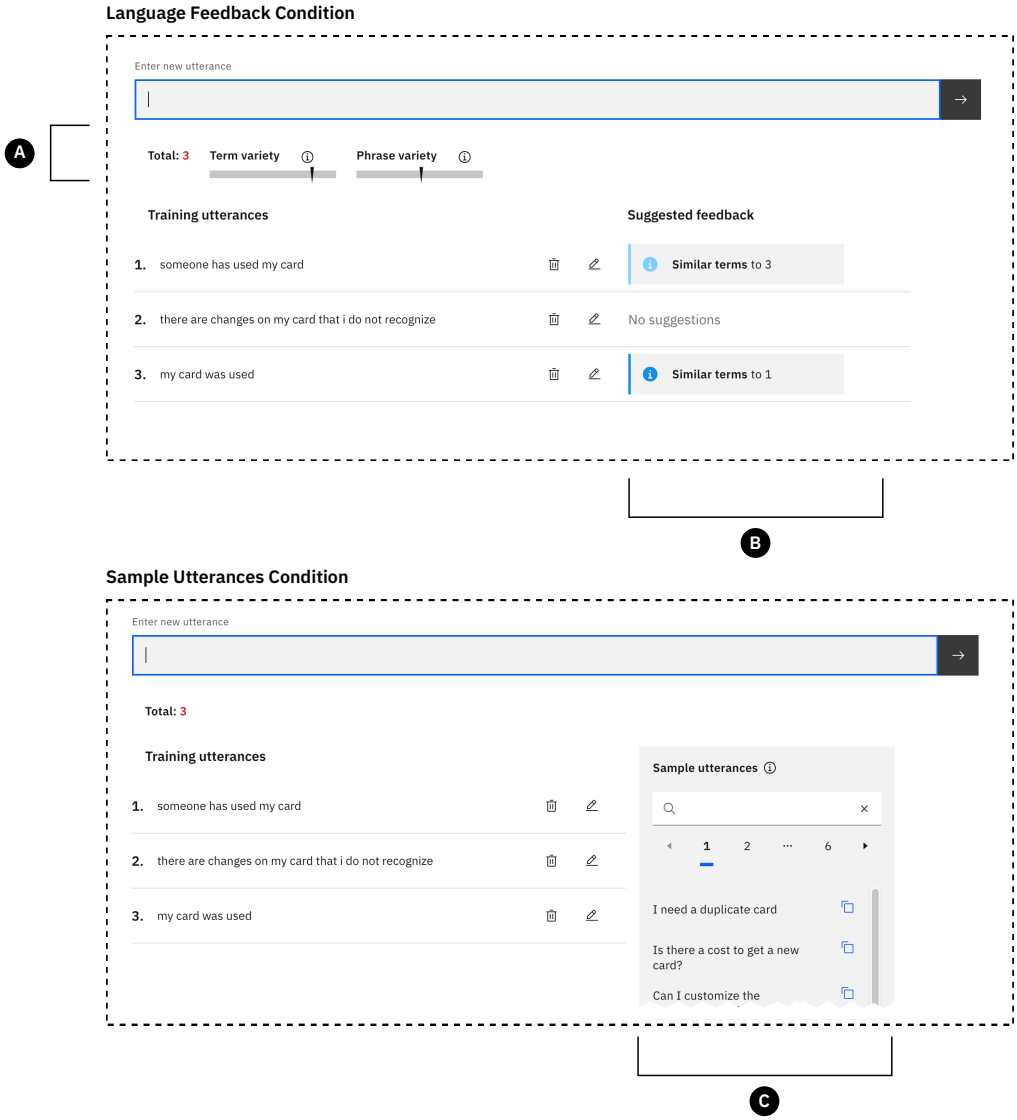

Figure 2: The utterance editor (See Figure 1-B) changed based on experimental condition. Top: The Language Feedback condition provided overall utterance variety feedback (A) and per-utterance feedback (B). Bottom: The Sample Utterances condition provided a searchable widget of sample utterances (C) that participants could copy by clicking on the icon.

4.4 Participants

We recruited participants from a multi-national technology company. Participants were recruited on internal Slack channels advertising the study and were compensated the equivalent of $12.50. The experiment took approximately 30 minutes to complete. All participants provided written informed consent and were treated in accordance with the guidelines for ethical treatment of human participants. We removed participants who did not complete both the tool and survey portions of the study, those who provided utterances that were clearly invalid (e.g. unrelated to the topic), and those who took excessively long breaks in the middle of the study (greater than one hour). Other than these performance-related exclusion criteria, there were no additional inclusion or exclusion criteria for participation. This resulted in 187 included participants, with 94 in the *Language Feedback* condition and 93 in the *Sample Utterance* condition. We will refer to participant numbers with "-L" (Language Feedback) or "-S" (Sample Utterances) suffixes to indicate their guidance condition. In line with our desire to target novice utterance writers, the majority of participants in each condition had little to no experience training conversational agents. 45% (*Language Feedback*) and 58% (*Sample Utterance*) of participants had never trained a conversational agent, 22% (*Language Feedback*) and 19% *Sample Utterance* had only done so a couple of times, and

Table 2: The types of language feedback supported by the language feedback condition in the tool. Note that "Term Variety" and "Phrase Variety" applied to all utterances as a whole, whereas the others were notifications that appeared (or did not) at the level of an individual utterance.

Feedback Type	Description
Term variety	Overall variety in terms across utterances.
Phrase variety	Overall variety of sentence structure across utterances.
Similar terms	Utterance has similar terms as other utterances. For example, multiple utterances contain "product" and "damage".
Similar phrasing	Utterance has similar sentence structure as other utterances. For example, "I need to return my product" and "I want to replace my order".
Filler words	Utterance has filler words (e.g. "okay", "please", "thanks").
Multiple topics possible	Utterance may contain multiple topics.
Long utterance	Utterance exceeds the recommended length, which was set at 17 words. This cutoff was determined by selecting the 85th percentile of utterance length in the Banking77 dataset.

Table 3: Participant demographics by study group, gender and age.

Condition - Gender	18-25	26-35	36-45	46-55	56-65	66-75	Total
Language Feedback - Male	10	8	3	9	6	0	36
Language Feedback - Female	8	19	15	6	10	0	58
Language Feedback - Prefer not to say	0	0	0	0	0	0	0
Sample Utterance - Male	5	11	15	4	4	0	39
Sample Utterance - Female	11	12	8	10	3	4	48
Sample Utterance - Prefer not to say	1	1	3	1	0	0	6

only 15% (*Language Feedback*) and 17% (*Sample Utterance*) do so once a month or more frequently. The distribution of those with experience with training any machine learning model was similar: 55% (*Language Feedback*) and 53% (*Sample Utterance*) had never done so, 20% (*Language Feedback*) and 16% (*Sample Utterance*) had only done so a couple of times ever and 12% (*Language Feedback*) and 17% (*Sample Utterance*) do so once a month or more frequently. Detailed demographics for participants are shown in Table 3 and their experience with conversational agents and machine learning are shown in Figure 3.

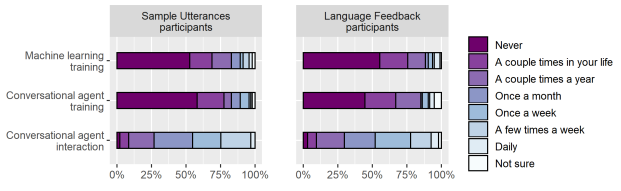

Figure 3: Participants' self-reported experience with conversational agents and machine learning: (1) how frequently they work on building or training machine learning models, (2) how frequently they work on building or training conversational agents, and (3) how often they interact with conversational agents in their daily life.

5 RESULTS

5.1 RQ1: Intent Recognition Accuracy

To evaluate the quality of the training utterances submitted by participants, we trained four classifiers: Language Feedback-Baseline, Language Feedback-Experimental, Sample Utterance-Baseline, and Sample Utterance-Experimental. Each classifier had four class intents, corresponding to the four prompts that were selected for the experiment.

We built intent classifiers using the *scikit-learn* library. Each classifier was trained using the approximately 960 utterances submitted by all the participants for that experimental condition (see Table 4 for exact utterance counts). To prepare the data for training, the utterances were transformed using the *all-MiniLM-L6-v2* sentence transformer and the labels were encoded using scikit-learn's label encoder. Then, the classifiers were built by training each on their corresponding data sets. For this we used scikit-learn's multinomial logistic regression model with all default settings including ridge regression (L2 regularization) for error reduction, class weights all being one, and no random state. Evaluations of the data were limited to this one model and default configuration as our concern was the *comparative* performance of the study data compared with the benchmark, and not on the model itself. Finally, the classifiers were evaluated on each of the four study intents using utterances from *Banking77* as the test set, which were intended to approximate end-user input.

Using the same configurations of the evaluation of participant utterances, we built a classifier consisting of all of the available training utterances from the *Banking77* set, limited to the four study

Table 4: Number of utterances used to train each classifier.

Classifiers	Total	Card Arrival	Compromised Card	Lost or Stolen Card	Order Physical Card
Language Feedback-Baseline	980	231	270	234	245
Language Feedback-Experimental	954	283	234	214	223
Sample Utterance-Baseline	953	264	213	216	260
Sample Utterance-Experimental	949	192	266	260	231
Banking77 Set (study intents only)	441	153	86	82	120

intents to act as a benchmark for classifier performance and compare with the performance of an agent trained on the participant-written utterances. We expected this classifier to be a best case to strive for, as using a partition of *Banking77* as the training set had two important qualities: (1) it is considered a high-quality, representative dataset of customer service queries in the banking domain and (2) the training and test utterances are subsets of the same overall dataset and thus are likely to include similar linguistic characteristics such as dialectal phrases, utterance length, and word choice for specific intents. If the classifiers trained using our participant-written utterances performed at similar accuracy levels as the one trained on the *Banking77* dataset, there would be strong evidence for the success of our participant-written utterances as a method of training an intent classifier.

The classifier accuracy results are presented in Table 5. All four classifiers that were trained on participant data showed strong performance, with accuracy ranging from 88.1% to 92.5% across the four classifiers. An exact binomial test was conducted for each classifier to compare its performance against what would be expected by chance (i.e., 25% accuracy, given the equal weighting of each of the four classes in the test dataset). All four classifiers showed significantly higher accuracy than chance performance, as shown in Table 5. Confusion matrices for each classifier are presented in the Supplementary Materials.

This performance was in line with that of the benchmark classifier which was both trained and tested using data from *Banking77*, with an accuracy differential between 7.5 and 3.1 percentage points lower than the benchmark results. Considering the similarity and the probable same source of the training and test datasets in *Banking77*, as well as the collective utterance writing inexperience of the participants, it is notable that the classifier performance is so close. This suggests that both kinds of support are similarly helpful and can achieve a training set as effective as a curated, conversation log-based set.

We additionally built classifiers for each condition based on participants' expertise in training conversational agents. Participants were split into two groups: *novices* (157 participants who reported "A couple of times a year" or less frequently) and *experts* (23 participants who reported "Once a month" or more frequently). Participants who reported "Not sure" were not included. However, due to the lack of training data for experts affecting model performance, we were unable to analyze the data further based on expertise. (Model evaluation details are provided in supplementary materials.)

5.2 RQ2: How Guidance Affected User Experiences

Although the experimental conditions resulted in similar intent training outcomes as the baseline condition, participants' comments indicate that having access to guidance altered their utterance writing approaches and improved aspects of their user experience. In this section, we report on participant effort, strategies participants implemented, and their overall impressions across conditions.

5.2.1 Differences in effort. One facet of effort is the amount of time it took participants to complete the task. To quantify differences, we ran a two-tailed paired t-test for each guidance condition comparing the time to complete the task in the baseline condition vs. the experimental condition. Language Feedback participants took an average of 86.5 seconds *more* time in the experimental condition than in the baseline condition (mean: 413.6 vs. 327.1, $t = 4.0605$, $df = 80$, $p < 0.0001$). Sample Utterance participants took an average of 37.2 seconds *less* time in the experimental condition than in the baseline condition (mean: 323.0 vs. 320.2, $t = 2.0520$, $df = 82$, $p = 0.043$).[4] Figure 4 provides histograms for the total time spent on the task across conditions.

This difference in timing is supported by the participants' responses on the post-study questionnaire. Participants reported higher mental demand (mean: 3.1 vs. 2.3; $t = 3.0426$, $df = 185$, $p = 0.0027$) and higher effort (mean: 2.8 vs. 2.1; $t = 2.7342$, $df = 185$, $p = 0.0069$) in the Language Feedback condition compared to the Sample Utterance condition. None of the other dimensions such as frustration, complexity, perceived success, or others were significantly different between the two treatment conditions. It may seem that these results are explained by the Sample Utterances condition potentially reducing to an easy copy-and-paste task compared to the Language Feedback condition, but this was not the case. We detail why in the next section.

5.2.2 Differences in strategy. In the post-experiment survey, we asked participants what strategies they used to write utterances in each condition. To analyze the open-ended responses from our participants' survey data, we used a qualitative coding approach. First, two researchers iteratively defined the code set by sampling the data until no new codes emerged. Using the agreed upon set of codes, they independently coded the remaining data and calculated reliability at the end. This process was repeated for each of the three conditions: baseline (Language and Sample Utterance participants

[4]We chose to remove outliers from each condition using the common definition of any values that were outside 1.5 times the inter-quartile range of the first and third quartiles.

Table 5: The intent recognition results across all conditions. *p*-values are derived from an exact binomial test conducted for each classifier against chance performance of 25%.

Condition Group	Accuracy	Precision (weighted)	Recall (weighted)	F1 (weighted)	*p*-value
Language Feedback-Baseline	0.913	0.913	0.913	0.912	< .0001
Language Feedback-Experimental	0.900	0.903	0.900	0.899	< .0001
Sample Utterance-Baseline	0.881	0.882	0.881	0.880	< .0001
Sample Utterance-Experimental	0.925	0.927	0.925	0.925	< .0001
Banking77	0.956	0.957	0.956	0.956	

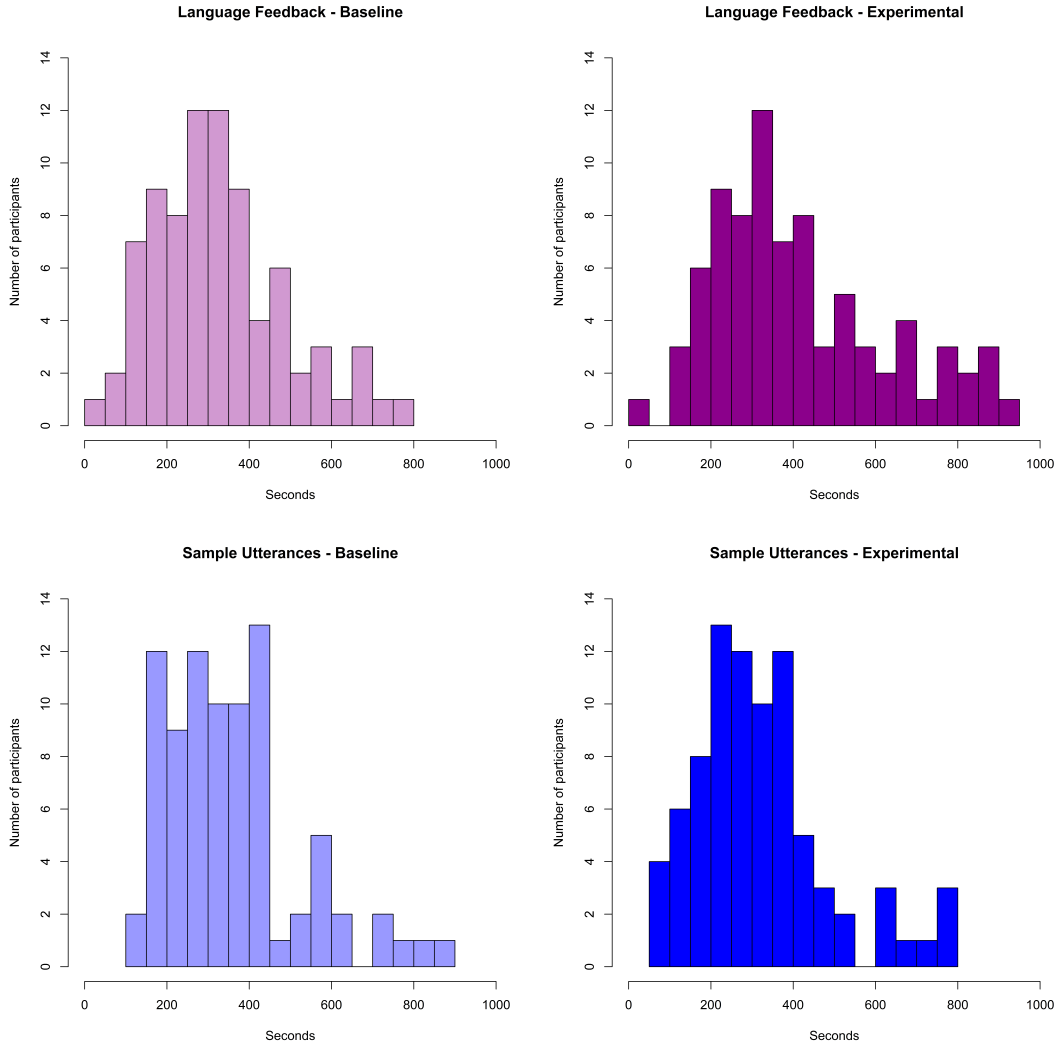

Figure 4: Histogram total task time grouped by experimental condition. The x-axis indicates the time to complete the task (providing utterances for one intent) in seconds. The y-axis is the count of participants. Each bucket is grouped in 50-second increments.

combined), Language Feedback-Experiment, and Sample Utterance-Experiment. All coded responses had high agreement with Jaccard indices ranging from 78% to 98% across the three qualitative coding efforts.

In the baseline condition, participants' strategies aligned with the dos and don'ts provided in the tool (Table 6). For example, to increase utterance variety, participants considered how they would speak with the agent in different scenarios, how different customers

Table 6: The strategies reported by the 187 participants in baseline condition.

Baseline Strategy	Definition	Count (%)
Self scenario	Participants considered how they would interact with a banking conversational agent (hypothetical) or how they've interacted in the past (experiential).	93 (49.7%)
Increase variety	Participants increased the variety of terms, phrasing, and types of utterances in response to instructions.	49 (26.2%)
Others' scenarios	Participants considered how others would interact with a banking conversational agent (hypothetical).	34 (18.2%)
Conciseness	Participants emphasized creating concise utterances.	11 (5.9%)
Intent description	Participants considered topics and content in the intent description and incorporated into their utterances.	9 (4.8%)
Natural human conversation	Participants modeled how different people speak or type (e.g. grammar, capitalization) to conversational agent in their utterances.	8 (4.3%)
Key words	Participants used key words as a starting point to write their utterances.	6 (3.2%)

would speak to the agent, or specific changes to increase sentence variety. About half of participants (93 of 187) considered what they would say in the given situation – for example, P27-L said, *"My general strategy was to write down the questions if I have faced a similar situation in my life and how I would seek help from a chatbot based on it."* Three participants varied utterances based on emotional state, such as P133-L, who said, *"I tried to have different feelings behind it such as patience, frustration, etc. to determine if I sent statements, questions, and their promptness,"* or P2-L who said, *"[My strategy was to] think of different situations like me being in a hurry, being stressed, etc."* 34 participants tried to put themselves in the shoes of others when generating example utterances, as captured by P23-S, who said, *"I tried to think of it as ... [how] someone like my parents who are not familiar with the online processes would ask."* About a quarter of participants (49) noted that they modified vocabulary or phrasing, but did not specify how. Additional strategies mentioned by participants also came from the list of dos and don'ts, with 11 participants mentioning building concise sentences and 8 participants focusing on reflecting natural speech. In short, for the baseline condition, participants did not stray far from the limited guidelines provided.

According to the post-experiment survey, 34 of 94 Language Feedback participants changed their strategies in their experiment conditions compared to the baseline. Table 7 details these changes. Comments from participants who changed strategies indicated that Language Feedback encouraged them to be more thoughtful in crafting diverse, high quality utterances: the newly reported strategies included an increased focus on variety (18 participants) and revisions to utterances in response to feedback provided by the tool (11 participants). Although only 11 participants reported reacting to feedback, the logs revealed that these participants made more edits in the experimental condition compared to the baseline. 58 participants made at least one edit in either Language Feedback-Baseline or Language Feedback-Experiment. In comparing these participants' editing in the experimental condition to the baseline, 50 participants made more edits, 5 made the same number of edits, and 3 made fewer edits. Participants averaged 3.2 edits during the experimental condition versus 0.7 edits in the baseline. Together with the previous findings regarding timing and effort, participants may have been more considerate in crafting their utterances as a result of Language Feedback, as confirmed by P7-L who said, *"As I created the utterances, I was reviewing the feedback and was determined to be more creative."*

The Sample Utterance condition had more participants report changes in strategy, with 44 out of 93 participants indicating new approaches. Table 8 reports these strategies. Participants' comments on these strategies indicated that sample utterances augmented their own utterances with new ideas and end user language. 13 participants reported a strategy of copying the provided samples. Log data reveals that 40 participants used the available samples, and either took the sample utterances as they were, or made additional changes prior to adding them. P97-S is one example of a "copy and modify" approach: *"[I utilized] some sample responses as a baseline for framing the utterances."* The other 53 participants chose not to copy utterances from the samples at all. Instead, they leveraged the samples as a source of inspiration for ideating new training utterances (21 participants), increasing variety (6 participants), and validating their own training utterances (2 participants). P96-S gave an example of how the tool supported their ideation: *" I read through what the real customers issues were and I found a couple that I wouldn't have thought of because I don't have that issue."* Sample utterances also served as a source for seeing how others interact with the conversational agent. P86-S said, *"[The sample utterances] reminded [me] that how I write is not how everyone writes (punctuation and sentence style) ... I've provided a mix of questions and statements, as well as sentence styles, phrases, and questions."* As an example of another strategy, P67-S used the samples as validation for training utterances they had crafted: *"Every time I checked to see if someone had wrote (sic) what I was thinking before I added my answer."* These participants viewed the sample utterances not as a source of ground truth to be blindly added for training, but as support for crafting their own utterances.

5.2.3 Usefulness of Language Feedback. Participants were asked to share what aspects of the Language Feedback condition they considered useful and not useful during the post-experiment survey. Their responses were coded using the same approach as how strategies were coded. Inter-rater reliability measured via Jaccard

Table 7: The updated strategies reported by the 34 out of 94 participants in Language Feedback condition.

Language Feedback Strategy	Definition	Count (%)
Increase variety	Participants increased the variety of terms, phrasing, and types of utterances in response to feedback.	18 (52.9%)
Respond to feedback	Participants modified their utterance in response to the specific feedback type received until they received no feedback.	11 (32.4%)

Table 8: The updated strategies reported by the 44 out of 93 participants in Sample Utterance condition.

Sample Utterance Strategy	Definition	Count (%)
Ideation	Participants used the the sample utterances to develop their own training utterances by searching through the feedback to get inspiration for new utterances.	21 (47.7%)
Copying samples	Participants reviewed the sample utterances and copied utterances that they considered high quality.	13 (29.5%)
Increased variety	Participants increased the variety of terms, phrasing, and types of utterances in response to their ability to view sample utterances	6 (13.6%)
Validation	Participants checked whether there was any alignment between their training utterance and the sample utterances to ensure they were on the right track.	2 (4.5%)
Overcome language barriers	Participants with language barriers (i.e., non-native English speakers) relied on sample utterances for writing assistance.	1 (2.3%)

Table 9: What participants found useful in the Language Feedback condition.

Useful	Description	Count (%)
Identifying similarity	Feedback helped participants identify similarities or duplicates that they had not previously noticed.	29 (30.8%)
Actionable guidance	Feedback helped participants identify specific issues within utterances and guided them to write higher quality utterances.	28 (29.8%)
Increased variety	Feedback helped participants feel like they were increasing the variety of terms, phrasing, and types of utterances.	16 (17.0%)
Increased thoughtfulness	Feedback helped participants consider their utterances more carefully and ensure best practices were followed.	11 (11.7%)

Table 10: What participants did not find useful in the Language Feedback condition.

Not Useful	Description	Count (%)
Similarity feedback type	Participants found the similar terms feedback type unhelpful because they felt it was either inaccurate (not true) or they were unable to diversify certain key words.	24 (25.3%)
Unclear or not actionable	Participants did not understand meaning behind feedback or how to respond to feedback.	15 (15.8%)
Minimal feedback received	Participant did not receive enough feedback to find the feature useful.	5 (5.3 %)

index indicated high agreement and measured 79% for what participants found useful and 93% for what participants found not useful. Results are presented in Tables 9 and 10.

Participants felt that Language Feedback was useful in the ways it was designed to be useful: by helping identify similarity (29 participants), providing actionable guidance (28 participants), and increasing variety (16 participants). 11 participants also reported being more thoughtful with crafting utterances. There was a marked increase in participants who modified their utterances in the Language Feedback condition, with 58 of 94 participants (61.7%) editing one or more utterances, up from 19 (20.2%) in the baseline condition, and 24 of 94 (25.5%) deleting one or more utterances, up from 7 (7.4%) compared to their baselines. In general, participants who preferred the Language Feedback condition over the baseline felt that responding to feedback notifications helped them write higher quality utterances.

Table 11: What participants found useful and not useful from the Sample Utterance condition.

Useful	Description	Count (%)
Spark new ideas	Sample utterances inspired participants to generate a new idea or topic they would not have thought of otherwise.	31 (32.6%)
Increase variety	Sample utterances helped participants feel like they increased their variety of terms, phrasing, and types of utterances used.	21 (22.1%)
Decrease cognitive effort	Sample utterances assisted participants develop utterances with less cognitive effort required.	10 (10.5%)
Validation	Sample utterances helped participants determine whether their training utterances were in accordance with best practice standards.	9 (9.5%)
Realistic	Sample utterances included real experiences in which participants had not encountered on their own, eliciting more diverse utterances.	8 (8.4%)

Table 12: What participants did not find useful from the Sample Utterance condition.

Not Useful	Description	Count (%)
Lower quality	Participants felt that the sample utterances were not relevant to the topic or lower quality than their own.	17 (17.9%)
Repetitive	Participants felt that there were many repeated sample utterances and they lacked variability.	15 (15.8%)
Inhibit creativity	Participants believed that having access to the sample utterances caused bias, creating less authentic or varied utterances.	11 (11.6%)
Excessive amount	Participants felt that there were too many sample utterances for the feature to be helpful.	7 (7.4%)

Participants' comments also suggested ways to improve the actionability and usefulness of language feedback. A quarter of participants (24) found feedback on similar terms or phrasing unhelpful. They felt that the notifications were flagging utterances that they did not consider similar or could not change without losing the intent. Since the term similarity feedback in the tool was agnostic to the banking topic, common topical words such as "card" could be highlighted, which participants thought should not be flagged. Additionally, 15 participants described ways the feedback fell short in explanations, either because it was unclear why that feedback was relevant or what to do about it. Participants described some of the feedback as *"vague"* (P81-L), *"not clear"* (P186-L), or that they *"did not understand [it]"* (P178-L). Additionally, participants found it unclear whether they were required to respond to the feedback notifications. 36 of the 94 (38.2%) feedback condition participants did not modify their utterances and 70 of 94 (74.4%) did not delete any utterances. More justification for feedback appearing and what to do about it was expected from these participants. Some participants proposed providing actionable suggestions in conjunction with feedback notifications (e.g. suggested synonyms to replace similar terms with), including examples, and implementing a scoring system to indicate when an optimal utterance set was achieved. Finally, five participants commented that they did not receive enough feedback, perhaps because the tool showed no indicators for high quality utterances beyond not providing feedback notifications.

5.2.4 Usefulness of Sample Utterances. Participants' responses on the usefulness of Sample Utterances were similarly qualitatively coded with an inter-rater reliability of 91% for what participants found useful and 95% for what participants did not find useful. Results are presented in Tables 11 and 12.

Similar to the Language Feedback condition, participants in the Sample Utterance condition found that the guidance supported ideation (31 participants), increased the variety of their training utterances (21 participants), and served as a way to check if they were on the right track (9 participants). Contrary to the increased thoughtfulness reported in Language Feedback, participants reported lower cognitive effort (10 participants) as noted by P148-S: *"It took a lot less mental effort to complete the task when [sample utterances] were there."* Participants also felt encouraged that the sample utterances represented realistic experiences (8 participants), especially noting that they helped improve breadth by *"[providing] the majority of main questions/situations that is associated with the situation"* (P94-S). Overall, participants who preferred the Sample Utterance condition over the baseline condition felt that the samples gave them new ideas when they got stuck and helped them write training utterances representative of end user language and scenarios.

Participants' responses on what was not useful indicate a need for better baselines over the sample utterances. We chose to provide all of the crowdsourced utterances as to not influence which ones participants selected, but as a result, participants in this condition reported a perception that sample utterances were lower quality (17 participants), which may have been exacerbated by repetition (15 participants) and the sheer amount (7 participants) made available. P163-S was frustrated by both the quality and quantity: *"[There are] too many of them to choose from. In parts, there were spelling and grammar mistakes."* 11 of the participants felt that having sample utterances limited their ability to develop their own, authentic

training utterances. Accordingly, we measured a significant difference in reported frustration from the post-experiment survey, with participants reporting a slight increase in frustration in the Sample Utterance condition (mean = 2.04) over the baseline (mean = 1.82) via a two-tailed paired t-test (t = 2.0452, df = 93, p = 0.0437). For example, P65-S noted that having the sample utterances available caused them to *"maybe [lose] the energy to think about my own ideas."* P46-S noted that having samples *"made me lazier in terms of thinking of other responses."* Participants proposed several enhancements to improve usefulness: an intuitive sample utterance panel that automatically hides incorporated utterances, AI features to sort and progressively disclose utterances as participants add their own, and additional search and filter capabilities.

6 DISCUSSION

6.1 Synthesis of Results

We designed the study to compare two new forms of guidance to support practitioners in training conversational agents: one where best practices for utterances are embedded into the tooling as interactive feedback (Language Feedback), and one where user-generated utterances are made available during the initial training phase (Sample Utterance). Instead, we found that participants in our study were generally good at following best practices, even without guidance, as shown in the baseline condition. Furthermore, perhaps due to the familiarity of the four intents from the banking domain, the overall accuracy of all the classifiers was high, making it difficult to measure the impact of either type of guidance on the classifier training (further discussed in Section 6.4). However, additional analysis revealed differences in how participants approached the task with different kinds of guidance and how those approaches influenced their abilities and perceptions when crafting training utterances.

When viewed holistically, the Language Feedback findings suggest that participants crafted well-considered training utterances. They took longer to create their utterances, made more edits, and reported higher difficulty and mental load. Approximately a third of participants reported changing utterances specifically in response to the feedback given, and over half mentioned an increasing variety in response to the feedback. Participants discussed how the feedback helped them find unwanted repetition and drew attention to where utterances could be improved. In short, the feedback in this condition was successful into guiding participants to follow the best practices as designed in the tool. Given this success, we posit that interactive, writing-time feedback can help novice utterance writers learn best practices more effectively than reading static documentation, helping to address the pain point of a steep learning curve for novices identified in the exploratory interviews. However, since best practices can vary between conversational agents, these types of language feedback should be customizable to a team's needs. Future work should also consider how to balance the trade-off between more thoughtful training utterances and increased task difficulty for users. For example, feedback could be surfaced for novice utterance writers as a means of learning, then eased out as they become proficient in writing high quality training utterances on their own.

In contrast, the Sample Utterance participants took less time, and reported less difficulty and mental load than their Language Feedback counterparts. We recorded fewer edits alongside the adoption of the provided sample utterances. Surprisingly, the majority of participants decided against copying and pasting the sample utterances at all, instead using the list to support ideation or validate existing ideas that they had. These participants leveraged the sample utterances as a resource to augment their work instead of doing their work for them. Given these uses and utterance writers' pain points of lacking access to user-generated examples (as identified in the exploratory interviews), we posit that gathering sample utterances in the initial training phase may be a valuable addition to support conversational agent developers. We found crowdsourcing to be a viable approach in this study; other methods such as leveraging quotes from customer support logs [14, 22, 25, 41] or using natural language generation may also provide similarly valuable user utterances without having to first deploy a conversational agent [10, 29].

6.2 Implications for Design

Participants reacted well to aspects of the guidance provided in both conditions, yet neither condition fully satisfied participants when crafting training utterances. In the Language Feedback condition, participants appreciated knowing when they were repeating words and phrasing, but also expressed frustration that the feedback was not always appropriate, could be confusing, and stifled their creativity by reducing their work to what amounted to a to-do list. In the Sample Utterance condition, most participants leveraged the many samples for inspiration and validation, but about a tenth of participants reported not thinking too hard about the task, perhaps aiming to finish it as quickly as possible. An improvement to our approach would be to design the tool to have the benefits from each approach. Our findings suggest that a tool should provide a balance between providing guidance, supporting ideation, and identifying appropriate candidate user utterances to reference.

Providing understandable, relevant and timely guidance. One limitation of the way the tool provided language feedback was the prominent and interrupting nature of the notifications, resulting in some participants treating them as a to-do list. A negotiated interruption style of feedback in which a user has control over when to be alerted may be more appropriate [26] and is preferred in the related domain of end-user debugging [37]. A useful framework that strikes a balance between being relevant and not distracting could be a surprise-explain-reward strategy [50] where notifications are subtle, but leave users satisfied with high quality explanations that are rewarding when engaged with. This framework helps explain why some participants did not understand the benefit of the feedback, as our tool fell short in explaining the impact of changes made by participants, which our exploratory interview participants also identified as a pain point. Ideally, future guidance should consider estimating the impact of changes made by intent trainers on the classifier. This remains a challenge as research has yet to address how to make accurate estimates without the high cost and time required for training.

Supporting ideation. Participants naturally leveraged the sample utterances to develop new ideas and validate existing ideas, similar to how examples have been shown to help (or hinder) ideation elsewhere [11, 42]. The strategies participants followed suggested that the list of sample utterances can be improved beyond the simple search functionality. For example, clustering algorithms can be used to group similar utterances together, and NLU techniques can be used to extract a summary of each cluster. Such approaches can help participants more quickly identify potentially relevant utterances based on their ideation or validation needs. Additionally, providing additional sources of sample utterances may inspire intent trainers in new ways. As mentioned earlier, natural language generation techniques and large language models may serve as additional starting points to augment end user language. Potential applications could be auto-completing example utterances or suggesting corrections as the intent trainers type their utterances. Simpler techniques such as providing synonyms and other related words may be similarly effective. Future work should investigate how different types of examples support the ideation of intent trainers.

6.3 Feedback Limitations

We designed the intent training tool to be sufficient to answer our research questions and consequently, the feedback mechanisms were intentionally simple. Knowledge of the domain or classifier to be trained can be leveraged to improve feedback. For example, to improve language feedback, we could augment the stop words based on the domain, or enable the intent trainer to flag the words most salient and ignoring those words when generating similarity feedback. Another limitation was that feedback was only given for the new intent being trained. In a real-world conversational agent, it would be useful to extend the feedback calculations to include intents that already exist and enable the trainer to potentially make changes to example utterances across intents. If such an agent has user utterance logs, those may serve as a basis to train a generative model to provide better auto complete suggestions or corrections to the intent trainer.

6.4 Threats to Validity

This study was conducted within one international technology company with its own tools, institutions, and culture that may not cleanly translate to other companies or institutions. Within the company, we recruited a variety of job roles and expertise, but approximately half reported experience with building conversational agents or other machine learning expertise. The exploratory interviews were similarly limited to participants of the same company and raise the same concerns about generalizability. The study limited its evaluation of the conversational agent to intent training and at a very small scale. Real-world conversational agents typically also have a dialogue component that dictates the conversational flow and mechanisms for detecting when conversations are going awry, both of which were out of scope in this experiment. A more complete evaluation would incorporate those aspects of conversational systems along with a more complete set of intents handled by the agent. Lastly, the intents in this experiment were intentionally chosen to be ones familiar to our expected participant pool to avoid the need for domain expertise. However, the choice may have been *too* common, as participants were able to craft high-quality utterances without much support, as indicated by the high classifier performance in the baseline condition. Future work should explore if such guidance would have similar impact for less easily identifiable intents.

7 CONCLUSION

In this paper, we evaluated how different types of guidance affect how participants approach and experience writing utterances for conversational agent intent training. We found that the type of guidance affected the time taken and effort required by participants, strategies for utterance writing, and their impressions of the task. Namely,

(1) Language Feedback participants took more time but were more intentional, and were also more likely to modify their utterances to adhere to best practices.
(2) Most Sample Utterance participants found creative uses for the samples, including validating their ideas or using them as inspiration for new training utterances, highlighting the importance of examples for ideation.
(3) In-situ assistance provided by intent training tools can be made more useful by providing context-appropriate and relevant guidance in a way that is controllable by the end user.

Incorporating a subset or combination of these forms of guidance could both facilitate the training of new conversational agents and also improve the experience of end users interacting with the agent.

REFERENCES

[1] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the People: The Role of Humans in Interactive Machine Learning. *AI Magazine* 35, 4 (2014), 105–120. https://doi.org/10.1609/aimag.v35i4.2513
[2] Ashley Belanger. [n. d.]. Air Canada must honor refund policy invented by airline's chatbot. *ArsTechnica* ([n. d.]). https://arstechnica.com/tech-policy/2024/02/air-canada-must-honor-refund-policy-invented-by-airlines-chatbot/
[3] Adam Benvie, Eric Wayne, and Arnold. Matthew. 2020. *Watson Assistant Continuous Improvement Best Practices.* https://www.ibm.com/downloads/cas/V0XQ0ZRE Accessed: 2023-09-04.
[4] Avrim L Blum and Pat Langley. 1997. Selection of relevant features and examples in machine learning. *Artificial intelligence* 97, 1-2 (1997), 245–271.
[5] Heloisa Candello, Claudio Pinhanez, Michael Muller, and Mairieli Wessel. 2022. Unveiling Practices of Customer Service Content Curators of Conversational Agents. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–33. https://doi.org/10.1145/3555768
[6] Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient Intent Detection with Dual Sentence Encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI.* Association for Computational Linguistics, Online, 38–45. https://doi.org/10.18653/v1/2020.nlp4convai-1.5
[7] Youngjin Chae and Thomas Davidson. 2023. Large language models for text classification: From zero-shot learning to fine-tuning. *Open Science Foundation* (2023).
[8] Yun-Nung Chen, Dilek Hakanni-Tür, Gokhan Tur, Asli Celikyilmaz, Jianfeng Guo, and Li Deng. 2016. Syntax or semantics? knowledge-guided joint semantic frame parsing. In *2016 IEEE Spoken Language Technology Workshop (SLT).* IEEE, 348–355. https://doi.org/10.1109/SLT.2016.7846288
[9] David Coniam. 2014. The linguistic accuracy of chatbots: usability from an ESL perspective. *Text & Talk* 34, 5 (2014), 545–567. https://doi.org/10.1515/text-2014-0018
[10] Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Zihao Wu, Lin Zhao, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, et al. 2023. AugGPT: Leveraging ChatGPT for Text Data Augmentation. *arXiv preprint arXiv:2302.13007* (2023). https://doi.org/10.48550/arXiv.2302.13007
[11] Hicham Ezzat, Marine Agogué, Pascal Le Masson, Benoit Weil, and Mathieu Cassotti. 2020. Specificity and abstraction of examples: Opposite effects on fixation for creative ideation. *The Journal of Creative Behavior* 54, 1 (2020), 115–122.

[12] James Glass, Eugene Weinstein, Scott Cyphers, Joseph Polifroni, Grace Chung, and Mikio Nakano. 2005. A framework for developing conversational user interfaces. In *Computer-Aided Design of User Interfaces IV: Proceedings of the Fifth International Conference on Computer-Aided Design of User Interfaces CADUI'2004 Sponsored by ACM and jointly organised with the Eight ACM International Conference on Intelligent User Interfaces IUI'2004 13–16 January 2004, Funchal, Isle of Madeira*. Springer, 349–360.

[13] Ong Sing Goh, Cemal Ardil, Wilson Wong, and Chun Che Fung. 2007. A black-box approach for response quality evaluation of conversational agent systems. *International Journal of Computational Intelligence* 3, 3 (2007), 195–203.

[14] Anuj Goyal, Angeliki Metallinou, and Spyros Matsoukas. 2018. Fast and scalable expansion of natural language understanding functionality for intelligent agents. *arXiv preprint arXiv:1805.01542* (2018).

[15] Xu Han, Michelle Zhou, Matthew J Turner, and Tom Yeh. 2021. Designing effective interview chatbots: Automatic chatbot profiling and design suggestion generation for chatbot debugging. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[16] Braden Hancock, Antoine Bordes, Pierre-Emmanuel Mazare, and Jason Weston. 2019. Learning from dialogue after deployment: Feed yourself, chatbot! *arXiv preprint arXiv:1901.05415* (2019).

[17] Matthew Henderson, Iñigo Casanueva, Nikola Mrkšić, Pei-Hao Su, Tsung-Hsien Wen, and Ivan Vulić. 2019. ConveRT: Efficient and accurate conversational representations from transformers. *arXiv preprint arXiv:1911.03688* (2019).

[18] Fred Hohman, Kanit Wongsuphasawat, Mary Beth Kery, and Kayur Patel. 2020. Understanding and visualizing data iteration in machine learning. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.

[19] Junhan Kim, Jana Muhic, Lionel Peter Robert, and Sun Young Park. 2022. Designing chatbots with black americans with chronic conditions: Overcoming challenges against covid-19. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–17.

[20] Yen-Ting Lin, Alexandros Papangelis, Seokhwan Kim, Sungjin Lee, Devamanyu Hazarika, Mahdi Namazifar, Di Jin, Yang Liu, and Dilek Hakkani-Tur. 2023. Selective in-context data augmentation for intent detection using pointwise v-information. *arXiv preprint arXiv:2302.05096* (2023).

[21] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *Comput. Surveys* 55, 9 (2023), 1–35.

[22] Peng Liu, Dong Zhou, and Naijun Wu. 2007. VDBSCAN: varied density based spatial clustering of applications with noise. In *2007 International conference on service systems and service management*. IEEE, 1–4.

[23] Lefteris Loukas, Ilias Stogiannidis, Odysseas Diamantopoulos, Prodromos Malakasiotis, and Stavros Vassos. 2023. Making llms worth every penny: Resource-limited text classification in banking. In *Proceedings of the Fourth ACM International Conference on AI in Finance*. 392–400.

[24] Wookjae Maeng and Joonhwan Lee. 2021. Designing a chatbot for survivors of sexual violence: Exploratory study for hybrid approach combining rule-based chatbot and ml-based chatbot. In *Asian CHI Symposium 2021*. 160–166.

[25] Neil Mallinar, Abhishek Shah, Rajendra Ugrani, Ayush Gupta, Manikandan Gurusankar, Tin Kam Ho, Q Vera Liao, Yunfeng Zhang, Rachel KE Bellamy, Robert Yates, et al. 2019. Bootstrapping conversational agents with weak supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 9528–9533.

[26] Daniel C McFarlane. 2002. Comparison of four primary methods for coordinating the interruption of people in human-computer interaction. *Human-computer interaction* 17, 1 (2002), 63–139.

[27] M de O Meira and AM de P Canuto. 2015. Evaluation of emotional agents' architectures: an approach based on quality metrics and the influence of emotions on users. In *Proceedings of the world congress on engineering*, Vol. 1. WCE London, 1–8.

[28] Joonas Moilanen, Aku Visuri, Elina Kuosmanen, Andy Alorwu, and Simo Hosio. 2022. Designing personalities for mental health conversational agents. In *Joint Proceedings of the ACM IUI Workshops*.

[29] Anders Giovanni Møller, Jacob Aarup Dalsgaard, Arianna Pera, and Luca Maria Aiello. 2023. Is a prompt and a few samples all you need? Using GPT-4 for data augmentation in low-resource classification tasks. *arXiv preprint arXiv:2304.13861* (2023).

[30] Milad Moradi, Kathrin Blagec, Florian Haberl, and Matthias Samwald. 2021. Gpt-3 models are poor few-shot learners in the biomedical domain. *arXiv preprint arXiv:2109.02555* (2021).

[31] Kyle Orland. [n. d.]. NYC's government chatbot is lying about city laws and regulations. *ArsTechnica* ([n. d.]). https://arstechnica.com/ai/2024/03/nycs-government-chatbot-is-lying-about-city-laws-and-regulations/

[32] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35 (2022), 27730–27744.

[33] Sara Pérez-Soler, Sandra Juarez-Puerta, Esther Guerra, and Juan de Lara. 2021. Choosing a chatbot development tool. *IEEE Software* 38, 4 (2021), 94–103.

[34] Raul Puri and Bryan Catanzaro. 2019. Zero-shot text classification with generative language models. *arXiv preprint arXiv:1912.10165* (2019).

[35] Nicole M Radziwill and Morgan C Benton. 2017. Evaluating Quality of Chatbots and Intelligent Conversational Agents. *arXiv preprint arXiv:1704.04579* (2017). https://doi.org/10.48550/arXiv.1704.04579

[36] Leonardo Rigutini, Achille Globo, Marco Stefanelli, Andrea Zugarini, Sinan Gultekin, and Marco Ernandes. 2024. Performance, energy consumption, and costs: a comparative analysis of automatic text classification approaches in the Legal domain. *International Journal on Natural Language Computing (IJNLC)* 13, 1 (2024).

[37] TJ Robertson, Shrinu Prabhakararao, Margaret Burnett, Curtis Cook, Joseph R Ruthruff, Laura Beckwith, and Amit Phalgune. 2004. Impact of interruption style on end-user debugging. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 287–294.

[38] Daniel Schloß. 2023. Towards Designing a NLU Model Improvement System for Customer Service Chatbots. (2023).

[39] Joseph Seering, Michal Luria, Connie Ye, Geoff Kaufman, and Jessica Hammer. 2020. It takes a village: integrating an adaptive chatbot into an online gaming community. In *Proceedings of the 2020 chi conference on human factors in computing systems*. 1–13.

[40] Vidya Setlur and Melanie Tory. 2022. How do you converse with an analytical chatbot? revisiting gricean maxims for designing analytical conversational behavior. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–17.

[41] Chen Shi, Qi Chen, Lei Sha, Sujian Li, Xu Sun, Houfeng Wang, and Lintao Zhang. 2018. Auto-dialabel: Labeling dialogue data with unsupervised learning. In *Proceedings of the 2018 conference on empirical methods in natural language processing*. 684–689.

[42] Pao Siangliulue, Kenneth C Arnold, Krzysztof Z Gajos, and Steven P Dow. 2015. Toward collaborative ideation at scale: Leveraging ideas from others to generate more creative and diverse ideas. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. 937–945.

[43] Annetta Sillard. 2022. *Shaping conversations: Investigating how conversational agents are designed and developed*. Ph. D. Dissertation. Kth Royal Institute of Technology. https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-321528

[44] Zeerak Talat, Aurélie Névéol, Stella Biderman, Miruna Clinciu, Manan Dey, Shayne Longpre, Sasha Luccioni, Maraim Masoud, Margaret Mitchell, Dragomir Radev, et al. 2022. You reap what you sow: On the challenges of bias evaluation under multilingual settings. In *Proceedings of BigScience Episode# 5–Workshop on Challenges & Perspectives in Creating Large Language Models*. 26–41.

[45] Gokhan Tur, Dilek Hakkani-Tür, and Robert E Schapire. 2005. Combining active and semi-supervised learning for spoken language understanding. *Speech Communication* 45, 2 (2005), 171–186.

[46] Michael Vetter. 2002. Quality aspects of bots. *Software quality and software testing in internet times* (2002), 165–184.

[47] William Yang Wang, Dan Bohus, Ece Kamar, and Eric Horvitz. 2012. Crowdsourcing the acquisition of natural language corpora: Methods and observations. In *2012 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 73–78.

[48] Yue Wang, Lijun Wu, Juntao Li, Xiaobo Liang, and Min Zhang. 2023. Are the BERT family zero-shot learners? A study on their potential and limitations. *Artificial Intelligence* 322 (2023), 103953.

[49] Jason D Williams, Nobal B Niraula, Pradeep Dasigi, Aparna Lakshmiratan, Carlos Garcia Jurado Suarez, Mouni Reddy, and Geoff Zweig. 2015. Rapidly scaling dialog systems with interactive learning. *Natural language dialog systems and intelligent assistants* (2015), 1–13.

[50] Aaron Wilson, Margaret Burnett, Laura Beckwith, Orion Granatir, Ledah Casburn, Curtis Cook, Mike Durham, and Gregg Rothermel. 2003. Harnessing curiosity to increase correctness in end-user programming. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 305–312.

[51] Yuanhao Xiong, Wei-Cheng Chang, Cho-Jui Hsieh, Hsiang-Fu Yu, and Inderjit Dhillon. 2022. Extreme Zero-Shot Learning for Extreme Text Classification. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 5455–5468.

[52] Mohammadali Yaghoubzadeh, Boualem Benatallah, M Chai Barush, and Shayan Zamanirad. 2019. A study of incorrect paraphrases in crowdsourced user utterances. *NAACL'19* (2019).