# Does Collaborative Human–LM Dialogue Generation Help Information Extraction from Human–Human Dialogues?

**Bo-Ru Lu** ♠* **Nikita Haduong** ♠* **Chia-Hsuan Lee** ♠ **Zeqiu Wu** ♠ **Hao Cheng** ♡
**Paul Koester** ♥ **Jean Utke** ♥ **Tao Yu** ♣ **Noah A. Smith** ♠♢ **Mari Ostendorf** ♠
♠University of Washington ♡Microsoft Research ♥Allstate
♣University of Hong Kong ♢Allen Institute for AI
roylu@washington.edu qu@cs.washington.edu

## Abstract

The capabilities of pretrained language models (LMs) have opened opportunities to explore new application areas, but applications involving human-human interaction are limited by the fact that most data is protected from public release for privacy reasons. Problem-solving human-human dialogues in real applications can be much more complex than existing Wizard-of-Oz collections, preventing successful domain transfer. To support information extraction (IE) for a private call center dataset (AIC), we introduce a human-in-the-loop dialogue generation framework capable of synthesizing realistic dialogues. In IE experiments with AIC dialogues, we observe 25% relative improvement in $F_1$ after augmenting a small set of real human-human conversations with synthetic data. In controlled experiments, we compare training with our human-in-the-loop-synthesized data vs. fully automatically LM-generated data and find that collaborating humans adds value both in the generation and annotation stages. We release code and our synthetic dataset to illustrate the complexity of call center conversations and encourage development of complex dialogue datasets that are more representative of natural data.

## 1 Introduction

Rapid advances in natural language processing have driven interest in its use in a wide variety of domains. However, applications that involve human-human interaction, such as call center dialogues, have had limited success (Lam et al., 2019; Albrecht et al., 2021; Pęzik et al., 2022; Lu et al., 2022). One reason is that natural problem-solving dialogues are not typically publicly available for privacy reasons, restricting opportunities for researchers to explore methods in advancing applications for these domains. Further, annotating private datasets can be expensive because of the need for in-house expertise, so training resources are limited. In this paper, we introduce a method to fill the data gap using synthetic data generated by a collaborative human–language model framework. Specifically, we experiment with a task of extracting information from auto insurance call center dialogues, using public synthetic data to improve performance on a private dataset.

Many available dialogue datasets are designed for training *virtual agents*, collected using pairs of humans to perform a task (Budzianowski et al., 2018; Rastogi et al., 2020; Chen et al., 2021). Designing for human-machine interaction results in dialogues that lack the complexity of human-human dialogues. Additionally, human-only data collection can have limited content diversity, result in imbalanced training sets, and does not scale to more complex tasks, due to the high cost of employing domain experts (Gururangan et al., 2018; Geva et al., 2019; Qian et al., 2021).

To reduce data collection costs, researchers have explored the use of language models (LMs) to generate synthetic training data (Liu et al., 2022a; Wang et al., 2023; Bao et al., 2023; Li

*Equal contribution. Code & data are available at https://boru-roylu.github.io/DialGen/.

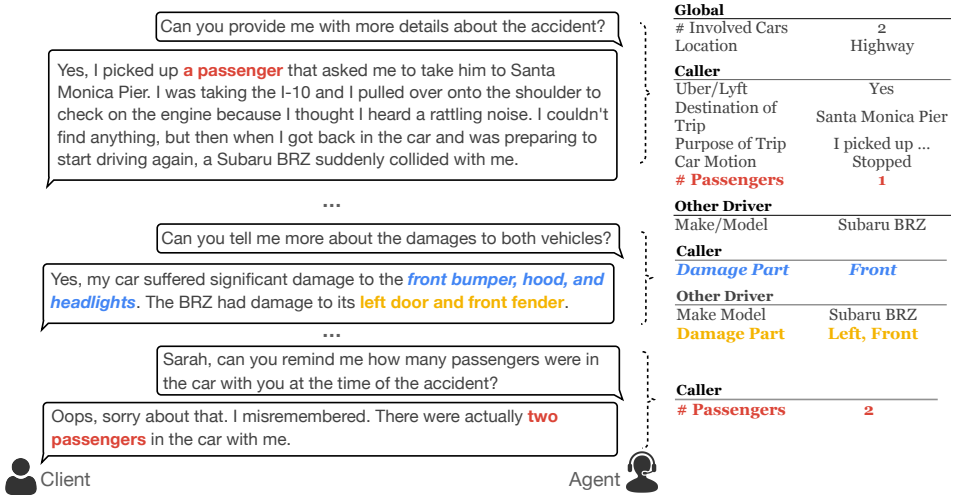


Figure 1: An illustrative snippet of our dialogue with entity-slot-value triples. **Yellow** is the slot with multiple values. ***Italic blue*** and **yellow** are the same slot (*Damage Part*) with different entities (e.g., *Caller* and *Other Driver*). **Red** is a slot with a value update.

et al., 2023; Asai et al., 2024). Synthesized data can target long-tail phenomena (Chu et al., 2020; Zhao et al., 2023; Zhang et al., 2023; Henning et al., 2023) and allow for public release of data that closely emulates real-world privacy-constrained domains, such as the medical domain (Park et al., 2018). Although LMs can follow instructions to generate text that closely resembles human writing, there can be challenges to ensure that the data are diverse and not too simplistic (Stahlberg & Kumar, 2021; Liu et al., 2022a). In addition, they still suffer from incoherence and consistency issues (Clark et al., 2021; Dou et al., 2022). To mitigate the shortcomings of LMs, human-LM collaboration can offer a robust solution, leveraging the strengths of both humans and machines (Sharma et al., 2022; Uchendu et al., 2023).

The idea of integrating human intelligence with artificial intelligence was initially introduced in Licklider (1960). Recent research has highlighted the proficiency of human-LM collaboration in generating a variety of data; however, most of the study focuses on short dialogues and text (Liu et al., 2022a; Bonaldi et al., 2022). In contrast, we investigate using a human-in-the-loop framework to create **lengthy and complex dialogues**.

Our work proposes a human-LM collaborative framework for dialogue generation (DIALGEN) that leverages the scalability and creativity of generative models, yet retains controllability through humans. Human collaborators edit the synthesized dialogues, which we use to boost information extraction performance on real-world call center data.

Many call center dialogues involve problem solving where customers provide information to an agent through question-answer pairs and clarifications that need to be interpreted in the context of the dialogue history. Our information extraction (IE) task is thus framed as an iterative information update after each agent-customer exchange, analogous to dialogue state tracking (DST) in task-oriented dialogues. However, unlike DST, the information extracted from each turn is collected to create a summary of the call rather than to generate a virtual agent's response or make an API call. In addition, the summary includes entities that are associated with attributes (slots) and values. To evaluate models on this IE task, we introduce entity-centric scoring methods that allow partial matching of multiple and descriptive values.

We demonstrate the effectiveness of DIALGEN by generating data in auto insurance calls, a domain with privacy restrictions that prevent public release of actual call recordings, and by performing information extraction. We work with a private dataset containing 34 dialogues with an average 197 utterances per dialogue and synthesize 235 dialogues with an average 46 utterances per dialogue. Experiments in our IE task show that additional synthetic data relatively improves model performance by 25% in the full $F_1$ score.

To summarize, our main contributions are:

- We design DIALGEN, a collaborative human-LM framework for generating complex dialogues in domains where privacy constraints have previously prevented data sharing with the research community. Synthetic data, training documentation and prompts will be released.
- We present DIALGEN-AIC, a custom dataset designed to illustrate the complexity of real-world auto insurance call center data. While not intended as a benchmark, DIALGEN-AIC aims to provide a demonstration of the complex nature of real conversations and the challenges faced in this domain, including linking information with different entities and tracking multiple values in a single slot.
- We propose an entity-centric scoring methodology that considers information links to different entities, allows for multiple slot values, and provides partial match scores for descriptive values.
- We compare our DIALGEN framework against a fully automatic LM framework and find that human collaboration adds value during both in generation and annotation.

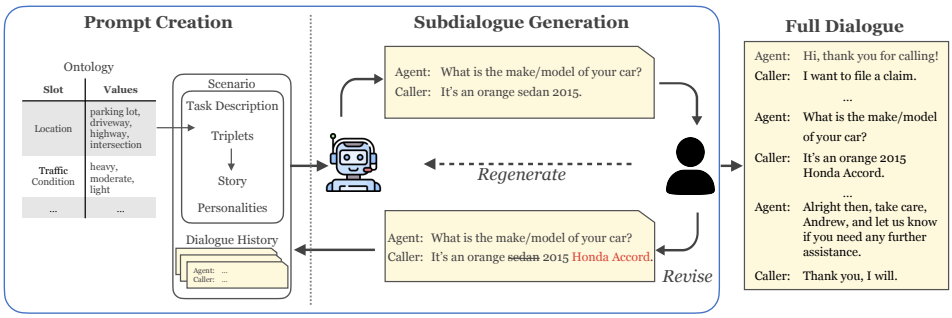


Figure 2: In the DIALGEN framework, a LM and a human reviewer collaborate to generate a dialogue. The LM creates subdialogues based on a scenario and dialogue history. The reviewer evaluates the subdialogue and can have the LM regenerate the subdialogue before revising it. The revised subdialogue is added to the dialogue history for generating the next subdialogue. This iterative process continues until the dialogue is complete.

## 2 Dialogue generation (DIALGEN)

As shown in Figure 2, our DIALGEN framework is designed to generate schema-guided dialogues through human-LM collaboration. An LM is selected as the backbone, then the data generation process begins with an initial task prompt consisting of natural language description for the desired dialogue (e.g., task description, desired slots, story, and personalities) and dialogue history. During each iteration, the LM first proposes a candidate subdialogue based on the history (the initial task prompt and the generated conversation so far). A human reviewer with sufficient domain knowledge then validate, edit and annotate the subdialogue, before requesting a continuation via an updated prompt to the LM. The reviewer can optionally augment the prompt with a specific instruction (see Table 8 in Appendix D) related to the desired dialogue flow. This process repeats until the dialogue is complete. At a high level, the human-in-the-loop mechanism ensures that the resulting dialogues are coherent and consistent with the prompt, covering desired content and fulfilling style specifications from domain experts. In the following, we describe each component of DIALGEN in detail.

### 2.1 Prompt for dialogue generation

The prompt for generating synthetic dialogues includes: the task description, entity-slot-value triplets, story, personality and dialogue history.[1]

[1]An example of a full prompt is given in Appendix D.1.

**Task Description.** Similar to the task descriptions given to humans in Wizard-of-Oz setups (Kelley, 1984), the template-based task description gives the information about the dialogue participants and the task scenario for the conversation, such as having the LM role-play as a user calling to file a claim with an agent at an insurance company, e.g., "*Role play car accident claim call. One person is Alice, an agent for a car insurance company, and the other is Bob, the caller who wants to file a claim.*"

**Entity-slot-value Triplets.** We randomly sample entity-slot-value triples from the expert-authored ontology to steer the LM to generate required content in the dialogue, allowing precise coverage of specific information, e.g., (*Caller, Injury, Neck*).

**Story.** Kim et al. (2022) synthesize social dialogues from triples of common sense knowledge by first using a social narrative to set up the scenario. We similarly use the randomly sampled triplets to generate a story with the LM before the dialogue generation. For example, the aforementioned entity-slot-value triple will be converted into the snippet of a story: "*The impact of the collision caused Bob's car to spin around and come to a stop. He immediately felt a sharp pain in his neck and knew that something was wrong*."

**Personality.** To enrich the diversity of callers, we randomly sample a personality from the list (Table 9 in Appendix D) for each dialogue, e.g., "*Bob is feeling distressed or frustrated due to the accident and its consequences.*" For the agent, we use the same personality for all dialogues, e.g., "*Alice is conversational, personable, patient, empathetic, sympathetic and professional*."

**Dialogue History.** The LM uses the full dialogue history to generate subdialogue that is consistent with the flow of the conversation. During the subdialogue generation, we append completed subdialogues before generating the next subdialogue. The initial dialogue history is always one exchange, e.g., "*Alice: Hi, thank you for calling DialGen Insurance! This is Alice. How may I help you today?*" followed by "*Bob: I am calling regarding a car accident*."

### 2.2 Human-in-the-loop subdialogue generation

The dialogue is generated iteratively where each subdialogue is revised by a human reviewer. Subdialogues are individually revised by a human trained to correct common LM errors such as those described by Dou et al. (2021), verify that required information is present (the sampled triples), and edit the text to meet stylistic criteria (e.g., adjusting tone). The reviewer can either revise individual turns directly or instruct the LM to regenerate specified turns, e.g., "*Have the caller correct earlier incorrect information*" (more examples in Table 8 in Appendix D). The LM may try to end the dialogue by including termination signals such as "*good bye*." If the LM ends the dialogue without covering the required triplets, the reviewer can delete and regenerate the turns.

### 2.3 Dialogue annotation

After a subdialogue is generated, a human annotator are asked to label spans in the dialogue that have information tuples associated with the task ontology. If a tuple in turn $t$ has a slot with the same referent and a different value than a previous turn, the human annotators are asked to resolve the duplication by indicating whether the new value is a correction UPDATE, KEEP, or additional detail to be concatenated with the previous value CONCAT. This annotation step is optional and can be decoupled from the framework depending on the target tasks or domains.

## 3 Problem definition and evaluation

An auto insurance call center dialogue involves a customer working together with an agent to address an issue or submit a claim. As the conversation progresses, the extracted information must be iteratively updated. This updating process is similar to the concept of dialogue state tracking (DST) used in task-oriented dialogues. However, unlike standard DST, the extracted information is used to summarize the call, not to make API calls or generate responses by a virtual agent.

### 3.1 Problem definition

Extracted structured information is typically represented as a collection of tuples $\{(s, v), s \in \mathcal{S}\}$, where $s$ is a slot label, $v$ is the associated value, and $\mathcal{S}$ is the full set of slots in the ontology. Values can be associated with a slot-dependent restricted set $\mathcal{V}_s$ or free-form text (e.g., a home address) or null. For multi-domain systems where different domains share some but not all slots (e.g., many domains have a date slot), the domain $d$ is separately tracked: $\{(d, s, v), d \in \mathcal{D}, s \in \mathcal{S}\}$. The full set of tuples is updated after each agent-user exchange to support construction of application calls needed to complete the task.

We formalize our information extraction task as follows. Ignoring domain for brevity, define $(A, U)_t$ as the pair of agent and user turns in exchange $t$. Given a sequence of exchanges between an agent and a user, $\{(A, U)_1, \ldots, (A, U)_t\}$, find the dialogue state $\{(s, v), s \in \mathcal{S}_t\}$, where $\mathcal{S}_t$ is the subset of slots active at time $t$ (i.e., having non-null values). The state associated with the final turn $T$ effectively provides a summary of the information extracted from the user in the dialogue.

### 3.2 Definition of extracted information

To accommodate the complexities of our dialogues, we augment the DST problem in three ways. First, we introduce the notion of a "referent", either with the global context or with the entity with which the extracted information is associated. Second, we allow slots to take on multiple values. Lastly, we allow slot values to be updated in multiple ways: a value can be corrected by the user, a new value can be added to form a list, or an existing value can be augmented, e.g., with details expanding on a free-form slot. Figure 1 provides an example of an agent gathering information about an accident together with the extracted tuples. There are three referents (*Global* context, *Caller*, and *Other Driver*); the number of passengers in the caller's vehicle was corrected from one to two; and the other driver's car has multiple *Damage Parts* (left and front).

With these changes, we describe our notation as follows, using the arrow diacritic to indicate cumulative state elements, upper case to indicate tuples and lower case to indicate labels or values, boldface to indicate a set of tuples, and calligraphic font to indicate a set of values. The initial dialogue state $\mathbf{X}_0$ is empty. The cumulative belief (CB) state $\overleftarrow{\mathbf{X}}_t$ (for $t > 0$) could be predicted directly or via a recursive state update: $\overleftarrow{\mathbf{X}}_t = \mathit{update}(\overleftarrow{\mathbf{X}}_{t-1}, \mathbf{X}_t)$, where only new/updated state values are predicted in the turn-level belief (TLB) $\mathbf{X}_t$ and the update function adds new slots and replaces updated slots. In the direct approach, it is possible to correct errors made by the model in previous turns, as well as introduce errors. A potential advantage of the update approach is that TLBs are shorter and therefore easier to predict.

Formally, $\overleftarrow{\mathbf{X}}_t$ and $\mathbf{X}_t$ are defined as follows. Define $\overleftarrow{\mathcal{R}}_t$ as the set of referents mentioned in a dialogue up through turn $t$, and $\mathcal{R}_t \subseteq \overleftarrow{\mathcal{R}}_t$ as the subset of referents associated with information updates in turn $t$.[2] The dialogue state and TLB after turn $t$, $\overleftarrow{\mathbf{X}}_t$ and $\mathbf{X}_t$, respectively, can both be represented as a set of referent-associated sets of active slots: $\overleftarrow{\mathbf{X}}_t = \{(r, \overleftarrow{\mathbf{S}}_{rt}), r \in \overleftarrow{\mathcal{R}}_t\}$ and $\mathbf{X}_t = \{(r, \mathbf{S}_{rt}), r \in \mathcal{R}_t\}$ where $\mathbf{S}_{rt} = \{S_{r1}, \ldots, S_{rn_{rt}}\}$, $n_{rt}$ is the number of active slots for referent $r$ updated at turn $t$, and $\overleftarrow{\mathbf{S}}_{rt}$ denotes the cumulative set of slots. An active slot is defined as $S_{rj} = (s_{rj}, \mathcal{V}_{rj})$, where $s_{rj} \in \mathcal{S}$ is the $j$th slot linked to the referent $r$, $\mathcal{S}$ is the set of slot (or domain-slot) types, and $\mathcal{V}_{rj}$ is a set of one or more values $v$ (categorical or free form text) associated with that slot. For our generated data, annotators are asked to provide state updates.

### 3.3 Evaluation

In IE tasks, precision, recall, and the F-measure are commonly used, while DST is based on joint goal accuracy (JGA) and slot accuracy.

[2]Our application uses a finite set of types $\overleftarrow{\mathcal{R}}_t \subseteq \mathcal{R}$, but it could be an open set, e.g., based on names.

Similar to DST, our IE task updates extracted information across turns. However, directly adopting DST metrics for dialogue-based IE is not ideal for two reasons. First, JGA is useful for DST because DST tasks require database queries that are built from the detected slots and values. Hence, accurate prediction is needed for all domains and slots, including null-valued instances. For a complex ontology, where many slots will be unfilled, JGA effectively emphasizes precision over recall. In contrast, for an IE task, the goal is to evaluate extraction quality, for which it is useful to look at precision/recall tradeoffs. Minor errors (e.g., an additional word in a non-categorical slot) should not significantly impact the readability of the extracted information. Second, in DST, queries are issued after most turns, so evaluating average performance at all turns makes sense. In contrast, in our IE task, information is accumulated (and corrected) for a final summary. In this case, turn averaging overemphasizes earlier parts of a conversation. For that reason, our IE metric evaluates the full state (CB) at specific dialogue points (quarter, half, three-quarters, end), and turn averaging is used for evaluating the prediction of state changes (TLB).

Our task requires the scoring to handle multi-value and extended free-form text responses. For scoring purposes, we treat multi-value slots as multiple instances of a slot. For free-form values, we adapt the multi-span setup in Li et al. (2022) and enumerate all possible alignments between the predicted and gold values. Each gold value is aligned to one predicted value at most, and percentage match is computed based on the longest common substring (LCS) to give a partial-credit score in $[0, 1]$ (rather than requiring exact match, i.e., $\{0, 1\}$ score) for use in measuring precision and recall.

**Cumulative Belief (CB) State Scores (evaluating $\overleftarrow{\mathbf{X}}$)** are computed for a particular turn (specific index $t$ or dialogue-final turn) in the $n$th dialogue, denoted as $m_{\text{CB}}(n, t) = \frac{1}{|\overleftarrow{\mathcal{R}}_{nt}|} \sum_{r \in \overleftarrow{\mathcal{R}}_{nt}} m(\hat{\overleftarrow{\mathbf{S}}}_{nrt}, \overleftarrow{\mathbf{S}}^*_{nrt})$, where $m$ can be precision ($P$) or recall ($R$). Overall scores are obtained by averaging over all dialogues $\mathcal{N}_t = \{n : \overleftarrow{\mathcal{R}}_{nt} \neq \emptyset\}$.[3] For example, precision is given by $\text{CB-}P(t) = \frac{1}{|\mathcal{N}_t|} \sum_{n \in \mathcal{N}_t} P_{\text{CB}}(n, t)$.

**Turn-level Belief (TLB) Scores (evaluating X)** are computed at the turn level, all of which are based on averaging over all $N$ dialogues in the test set formulated as $\frac{1}{N} \sum_n \frac{1}{|\mathcal{T}_n|} \sum_{t \in \mathcal{T}_n} m_{\text{TYPE}}(n, t)$, where $\mathcal{T}_n = \{t : \mathcal{R}_{nt} \neq \emptyset\}$ and $\text{TYPE} \in \{\text{TLB}, \text{R}, \text{RS}, \text{SV}\}$ denotes diagnostic score type. The scores ($m_{\text{TLB}}$, $m_{\text{R}}$, $m_{\text{RS}}$, $m_{\text{SV}}$) are described in Appendix A. For each turn, the $m_{\text{TLB}}$ is performance over the TLB; $m_{\text{R}}$ is how well referents are recognized; $m_{\text{RS}}$ is how well referents are associated with slots when ignoring values; and $m_{\text{SV}}$ is performance of slot-value detection when ignoring referents.

## 4 Datasets

We were provided with a private dataset of 34 natural auto insurance claim calls (AIC). In each call, the agent's task is to gather detailed information about an auto accident. The calls were human transcribed and labeled using a schema with six referents and sixty possible slots from ten domains (Appendix E.3). Calls had high variance in length and complexity, as shown in Table 5 in Appendix B. Additionally, 50% of dialogues had multiple values for at least one active slot. We split the calls into 7/4/23 for train/val./test sets aiming for a slot count split of 20/10/70.

Using AIC as a target dataset for augmentation, we apply DIALGEN with ChatGPT (`gpt-3.5-turbo-0301`) as the LM backbone to create DIALGEN-AIC, which contains 235 labeled dialogues (Appendix E.5). Reviewers completed a one-hour training to become familiar with the task and practiced generating one dialogue under supervision. Full training was complete after they received feedback for their first 3–5 dialogues. They were instructed to aim to generate dialogues with $\approx 50$ turns. On average, each dialogue comprises $8 \pm 4$ subdialogues, with 38% of turns receiving edits and 20% of turns being deleted. Each dialogue involves $9 \pm 10$ times of partial or full subdialogue regeneration.

[3] In the initial turns, there may be nothing to extract and no false predictions; thus $\overleftarrow{\mathcal{R}}_{nt} = \emptyset$.

| | AIC | DialGen AIC | AutoGen AIC |
|---|---|---|---|
| # dialogs | 34 | 235 | 195 |
| # turns / dialog | 197 | 46 | 31 |
| # tokens / dialog | 4195 | 1128 | 947 |
| # user tokens / turn | 18 | 22 | 27 |
| # agent tokens / turn | 25 | 27 | 35 |

Table 1: Statistics are calculated on the full dataset.

Data collection occurred over 2 months with multiple iterations as documentation and task instructions evolved to become more comprehensive and consistent. The final version of the task instructions further encouraged workers to update slot values in multiple ways and include multiple values in a slot (as described in §2.1). We follow the methodology in SQuAD (Rajpurkar et al., 2016), calculating inter-annotator agreement (IAA) at the turn level with three annotators and 32 dialogues, with a resulting IAA of 78.5% $F_1$ (Appendix E.2).

We investigate the importance of including human reviewers in DIALGEN by comparing fully automatically created dialogues with DIALGEN-AIC. These dialogues are generated by prompting ChatGPT using the same training scenarios. We refer to this LM-only framework as AUTOGEN and use it to generate the AUTOGEN-AIC dialogues. Prompt details are in Appendix D.2. In comparing AIC, DIALGEN-AIC, and AUTOGEN-AIC, we find that synthetic data has fewer turns, longer turns, and less variance in length, with fully automatic data being the most extreme. Adding a human in the loop results in much longer and more varied dialogues, but they are still far from the complexity of human-human dialogues (see Table 5 in Appendix B). Compared to MultiWOZ (Budzianowski et al., 2018), DIALGEN-AIC is more complex. MultiWOZ dialogues average 14 turns and 8 active slots per dialogue, compared to 46 turns and 38 slots on average for DIALGEN-AIC, and 198 turns and 48 slots for AIC. We split DIALGEN-AIC into train/val./test sets with a ratio of 80/10/10 dialogues, selecting val./test sets by randomly sampling from the final iteration of data collection. Table 1 contains additional statistics of AIC, DIALGEN-AIC and AUTOGEN-AIC.

## 5 Experiments

**In-context Learning.** Hu et al. (2022) propose IC-DST and use schema prompts and a specialized retriever to enable few-shot in-context learning to predict state change with an LM. Given longer dialogues, a more complex ontology, and more slots to track than the datasets, the representation of dialogue history becomes a crucial concern. The SQL tables of the ontology is 1696 tokens, and our chosen LM (`gpt-3.5-turbo-0301`) has a token limit of 4096 tokens. To accommodate the token constraints, we truncate the in-context examples when given a longer dialogue state. We extract the TLB at the turn $t$ and accumulate the TLBs as CB. Details of the SQL table are shown in Appendix D.3.

Furthermore, our task requires the model to identify the corresponding entity (referent) for the predicted slot-value pair. We redesign the prompt (Appendix D.3) to instruct the LM to generate the referent, slot, and value simultaneously. The retriever, SBERT (Reimers & Gurevych, 2019), is finetuned on the full DIALGEN-AIC training set, which is also used as the example selection pool. Due to privacy concerns, we only evaluate IC-DST on the DIALGEN-AIC test set.

**Finetuned Transformers.** We follow idea of the previous work (Lee et al., 2021; Lu et al., 2024) to independently extracted the information and finetune T5 (Raffel et al., 2020) and Long-T5 (Guo et al., 2022) with schema information embedded in the prompt. The models predict only active slots (together with referent and value) with one prompt per domain. The CB is the aggregate of predictions over all domains.

In addition, we explore four different configurations of prompt and model outputs:

**Long-T5†:** Use $\{(A, U)_\tau\}_{\tau=1}^{t-1}$ to predict CB.

| Method | $CB_{avg}$ | $CB_1$ | $CB_2$ | $CB_3$ | $CB_4$ | TLB |
|---|---|---|---|---|---|---|
| IC-DST | 71.3 | 71.9 | 68.5 | 68.4 | 68.2 | 68.1 |
| Long-T5† | 71.8 | 72.5 | 71.7 | 71.0 | 70.4 | – |
| Long-T5 | 66.3 | 64.3 | 64.8 | 64.3 | 63.9 | 68.5 |
| T5 | 76.8 | 78.4 | 74.9 | 73.7 | 74.1 | 73.9 |
| T5-SC | **78.2** | **79.3** | **76.4** | **76.6** | **76.9** | **74.2** |
| T5-SC§ | 78.5 | 78.7 | 76.2 | 76.0 | 76.2 | 75.0 |

Table 2: $F_1$ scores on the DIALGEN-AIC test set. § denotes results with name substitution.

**Long-T5:** Use $\{(A, U)_\tau\}_{\tau=1}^{t-1}$ to predict TLB; add to CB.

**T5:** Use $(A, U)_{t-1}$ to predict TLB; add to CB.

**T5-SC:** Use $(A, U)_{t-1}$ and previous domain CB to predict state change ΔCB; update CB.

Because the input length can be longer than 1k tokens, we choose Long-T5 to cover all turns with the prompt, while the T5-based models make predictions based on the current turn only. T5-SC further considers the state change ΔCB, which is similar to the TLB but augmented with the four state-change commands. Details of the finetuning and in-context learning configuration are given in Appendix C and details of the prompts for the different cases are given in Appendix D.4.

**Experimental Setup.** When conducting experiments involving AIC, the model selection criterion is the highest TLB $F_1$ score on the AIC validation set. For experiments solely on DIALGEN-AIC or AUTOGEN-AIC, models were chosen based on TLB $F_1$ score on the DIALGEN-AIC validation set. Additional hyperparameters can be found in Appendix C.1. All reported values represent the medians of five different random seeds.

## 6 Results

We report the main results on both cumulative and turn update scores. The cumulative scores are presented in two ways: $CB_{avg}$ as an average of CB across every user turn, and $CB_Q$ as the CB at user turn $t$, where $t = \lceil QT/4 \rceil, Q \in \{1, 2, 3, 4\}$ and $T$ is the total length of a dialogue. Thus, $t$ will be a specific turn, at either a quarter, a half, three-quarters, or the end of the dialogue. The score of the last cumulative belief state $CB_4$ is the full $F_1$ score and can be regarded as evaluating a conversation summary. Model development was done only on the synthetic data to minimize use of real data.

**Results on DIALGEN-AIC Test Set.** The results on DIALGEN-AIC with different learning strategies and T5 configurations are presented in Table 2. The performance of IC-DST is lower than all T5 variants, although this may be due to the difference in use of domain-specific prompts. Note that our IC-DST implementation is based on the same ChatGPT model used for generating the DIALGEN-AIC, so the low results suggest that human collaboration creates data sufficiently different from ChatGPT text such that ChatGPT cannot easily address this task. Predicting CB directly requires the full history, which is only possible with Long-T5. With Long-T5, there is a benefit to predicting CB directly over TLB. However, optimizations needed to handle a longer history have tradeoffs that result in performance that is worse than the standard T5 model with TLB prediction for this task. T5-SC achieves the best result, which updates values rather than adding them as new values in a list.

To mitigate the potential risk of LMs generating personal information linked to randomly generated names in shared data, we replace them with other randomly generated names. As shown in Table 2, T5-SC exhibits comparable performance on both the original and renamed dialogues, indicating that the renaming process does not impact the model's effectiveness.

**Results on AIC Test Set.** The two best models (T5 and T5-SC) are used in experiments on the real data (AIC). The $F_1$ results for different training sources are given in Table 3. The performance for the model trained on the synthetic data alone is better than with the small amount of the real data, but the best results are obtained by model trained on the combined

| Method | Data | $CB_{avg}$ | $CB_1$ | $CB_2$ | $CB_3$ | $CB_4$ | TLB |
|---|---|---|---|---|---|---|---|
| T5 | AIC | 38.3 | 39.6 | 37.1 | 36.2 | 35.1 | 34.8 |
| T5 | DIALGEN-AIC | 40.4 | 41.7 | 42.6 | 39.9 | 37.7 | 40.9 |
| T5 | AIC + DIALGEN-AIC | 43.7 | 42.9 | 42.2 | 43.0 | 41.9 | 43.7 |
| T5-SC | AIC | 39.2 | 40.0 | 38.1 | 37.1 | 36.1 | 33.9 |
| T5-SC | DIALGEN-AIC | 41.0 | 43.6 | 42.1 | 41.3 | 40.5 | 38.9 |
| T5-SC | AIC + DIALGEN-AIC | **46.2** | **47.8** | **47.2** | **45.9** | **45.3** | **44.6** |

Table 3: $F_1$ scores on the AIC test set for different training data.

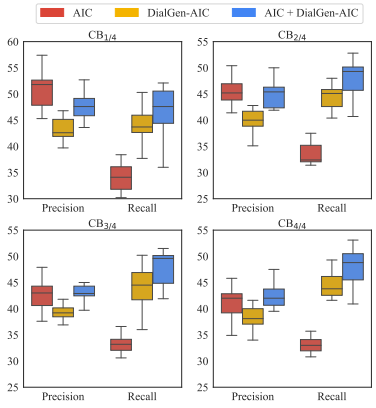


Figure 3: CB precision and recall scores on the AIC test set and T5-SC models.

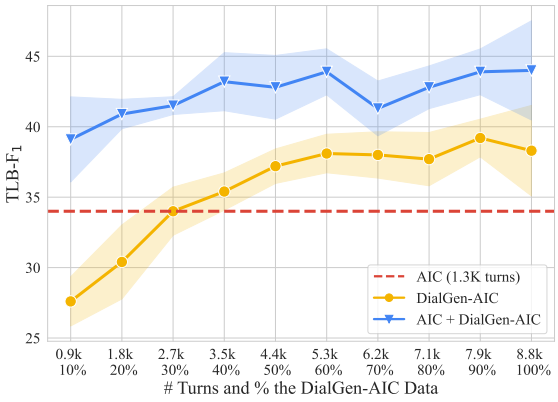


Figure 4: TLB scores for T5-SC on AIC test set by varying the amount of DIALGEN-AIC training data.

data. The difference between the best case model (T5-SC) on AIC alone vs. in combination with DIALGEN-AIC is significant with $p < .01$ using a bootstrap test. Because of the higher frequency of state changes in the human-human dialogues, there is a greater benefit from the T5-SC model for the real data, with an 8% improvement in the $CB_4$ score compared to 4% for the synthetic data when using all training data.

To provide more insight into performance, we present the precision/recall results for CB in Figure 3. Incorporating synthetic data yields higher recall and outperforms using real data alone in terms of $F_1$. The increased recall can be attributed to the inclusion of a wider range of values in the synthetic data, which are not covered by the AIC training set. However, this improvement comes at the expense of lower precision. By combining both data sets, the model achieves better alignment with real-world data while retaining the advantage of high recall scores from the synthetic data.

We also experimented with varying the amount of synthetic data used in training the model in order to determine the relative value of synthetic versus real data. Figure 4 shows that using 59 synthetic dialogues (approximately 2.7K turns) yields results similar to those obtained from the AIC training set, which consists of 1.3K turns in 7 dialogues. These results suggest that roughly 2.1 times as many turns of synthetic data is needed to match the performance of the real data, or 8.4 times as many synthetic dialogues since the synthetic dialogues are shorter. However, the synthetic data is more valuable in combination with real data, for which the benefit beyond 97 dialogues (50%) is minimal. This suggests an opportunity for further improvement through strategic scenario sampling.

**Comparison of DIALGEN with AUTOGEN.** To understand the importance of the role of humans in data synthesis, we investigate whether an LM annotator is sufficient and compare the value of dialogues generated by AUTOGEN and DIALGEN. First, we use ChatGPT and GPT-4 (OpenAI, 2024) to annotate DIALGEN-AIC training dialogues and compare the annotations with labels, obtaining 52.3 and 65.5 TLB accuracy respectively. This suggests GPT-4 is a better annotator than ChatGPT, which aligns with the findings of

| Framework | Dialogue Generator | Annotator | TLB |
|---|---|---|---|
| AUTOGEN | ChatGPT | ChatGPT | 49.5 |
| | ChatGPT | GPT-4 | 60.7 |
| DIALGEN | ChatGPT w/ Humans | GPT-4 | 63.2 |
| | ChatGPT w/ Humans | Humans | 73.9 |

Table 4: Comparison between different data synthesis frameworks. The TLB $F_1$ is obtained by finetuning a T5 models on generated data and testing on the DIALGEN-AIC test set.

Gray et al. (2023) and Tekumalla & Banda (2023). Then, we finetune T5 models on each setting and report TLB score to evaluate the quality of synthesized data (Table 4). T5 trained on dialogues generated by the DIALGEN framework and annotated by humans yielded the best result, demonstrating how a human-in-the-loop data synthesis framework can provide higher quality training data than a fully automated system.

### 6.1 Error analysis

Out of the 56 slots in the AIC test set, we noticed an improvement in 45 slots, while 4 slots were tied, and the remaining 7 slots have slightly worse performance. Our error analysis reveals two main categories for the performance loss: data mismatch between AIC and DIALGEN-AIC and over-reliance on surface-level features.

**Data Mismatch.** We lose performance for the slot *Car Mileage* because of a difference in language used when describing the mileage of a car. In AIC, agents ask a binary confirmation for whether the mileage on the vehicle is above a certain threshold, whereas callers in DIALGEN-AIC describe car mileage with an exact number. For the slot *Traffic Controls Obeyed*, AIC callers indirectly indicate that traffic controls are not obeyed, e.g. stating that the other driver ran a red light. In DIALGEN-AIC, the agent asks the caller to confirm directly whether traffic controls were obeyed.

**Surface Level Text.** The model both over- and under-predicts slots due to surface-level features such as predicting *Number of Involved Cars* when the text discusses counting vehicles, despite many such instances in AIC simply describing the traffic environment to contextualize the accident, e.g., there was a vehicle in front of the caller, but it was not involved in the accident. The model also predicted this slot when there was language about the number of passengers with a driver. Similarly, *Color* would be predicted whenever colors were mentioned, e.g., a purple bruise. *Traffic Flow* was severely under-predicted when it would have been beneficial for the model to predict the slot whenever it saw information describing lane direction.

## 7 Conclusion

We propose DIALGEN, in which humans and LMs collaborate to generate long, complex dialogues. We demonstrate its effectiveness by synthesizing auto insurance calls and conducting information extraction (IE) experiments. While we build on the DST framework, our IE experiments target an ontology and data that are more complex than the DST task was originally designed for. To serve the IE task, we introduce an entity-centric scoring methodology more suitable for our IE task than the conventional joint goal accuracy metrics used in DST. In our controlled experiments, we contrast the outcomes of training with data synthesized through a human-in-the-loop method against data generated and annotated entirely by ChatGPT or GPT-4. Our findings indicate that human-LM collaboration enhances the process at both the data generation and annotation phases, validating the importance of including humans in the data synthesis process to generate more realistic dialogues. Our experiments demonstrate that the data generated by DIALGEN, despite dissimilarities with the data it is designed to emulate, can significantly improve model performance for information extraction on real-world human dialogues.

## 8 Limitations

While DIALGEN can be used to generate synthetic data for privacy-constrained settings, the effectiveness largely depends on the LM employed, target setting, and language. We conducted all experiments in the auto insurance claim calls domain in English, where English is a high-resource language, and descriptions of car accidents are reasonably frequent in online text. An LM without reasonable capability in generating text in the target domain and language will result in low quality subdialogues, which can result in a frustrating collaboration for the human reviewers.

While DIALGEN results in dialogues that are more similar to real human-human interactions than AUTOGEN, our analysis shows that there are still substantial differences. This gap may be reduced by incorporating other forms of human feedback, more powerful LMs, and/or changing the prompt configuration. For example, subdialogue generation in DIALGEN is guided by including the full dialogue history as context for each subsequent subdialogue. LMs have finite context input length, so the max length of a generated dialogue is limited by the chosen LM. Methods to overcome this limitation can include truncating the dialogue history context, investigating which parts of the prompt contribute little to guiding the LM, and representing dialogue history in a more efficient manner.

In our controlled experiments comparing AUTOGEN and DIALGEN, the dialogue generation and annotation process in AUTOGEN is accomplished by a single LM agent with different associated prompts. There is room to improve AUTOGEN, such as exploring different prompt designs and introducing quality control agents.

## 9 Ethical considerations

Preserving privacy (Xin et al., 2020; Liu et al., 2022b; Torfi et al., 2022) is an important challenge in synthetic data generation. Ensuring important characteristics in synthesized data with DIALGEN requires a domain expert who may have access to real, private data and can unintentionally leak information. DIALGEN-AIC, on the other hand, generates personal information using the Faker package,[4] but there is a potential for the LM to produce personal details related to randomly created names. To mitigate the potential risk in shared data, we use gender guesser package [5] to detect the gender of each name and replace it with other same-gender name. If DIALGEN users plan to publicly release their data, they should remove potentially identifying information such as names from the synthesized data. In the released DIALGEN-AIC, we replace names with random alternatives to prevent the inadvertent generation of sensitive personal information by the LM.

Other than privacy issues, LMs can produce harmful content, and the risks of such production can increase depending on the target data setting. When employing humans to collaborate with LMs, practitioners should determine whether additional safety features such as toxic language filters are required to protect the workers.

Regarding the data collection hiring process, all dialogue reviewers were recruited from university listings and compensated at a rate of $18.69 per hour, following university practices. Prior to data collection, we instructed our reviewers to familiarize them with the ontology, annotation guidelines, and criteria for assessing dialogue quality. We established a Slack workspace for smooth communication with the workers throughout the process, providing feedback and promptly addressing questions and concerns they raised. This interaction ensured high quality of the gathered data.

## Acknowledgments

We would like to express our sincere gratitude to Kevin Everson, Yanda Chen, and Yushi Hu for their invaluable discussions and preliminary studies. We would also like to thank Bing-

[4] https://github.com/joke2k/faker

[5] https://github.com/lead-ratings/gender-guesser

Syuan Wang and Irene Wang for their expert web programming consulting and debugging support. Additionally, we extend our appreciation to members of UWNLP for their valuable insights and contributions throughout the project. Lastly, we are grateful to the diligent student reviewers from the University of Washington for their dedicated efforts in data creation. Their contributions were essential to the success of this research.

## References

Tobias Albrecht, Theresa Maria Rausch, and Nicholas Daniel Derra. Call me maybe: Methods and practical implementation of artificial intelligence in call center arrivals' forecasting. *Journal of Business Research*, 2021. URL https://api.semanticscholar.org/CorpusID:225140800.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=hSyW5go0v8.

Jianzhu Bao, Rui Wang, Yasheng Wang, Aixin Sun, Yitong Li, Fei Mi, and Ruifeng Xu. A synthetic data generation framework for grounded dialogues. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10866–10882, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.608. URL https://aclanthology.org/2023.acl-long.608.

Helena Bonaldi, Sara Dellantonio, Serra Sinem Tekiroğlu, and Marco Guerini. Human-machine collaboration approaches to build a dialogue dataset for hate speech countering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 8031–8049, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.549.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 5016–5026, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1547. URL https://aclanthology.org/D18-1547.

Derek Chen, Howard Chen, Yi Yang, Alexander Lin, and Zhou Yu. Action-based conversations dataset: A corpus for building more in-depth task-oriented dialogue systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3002–3017, 2021.

Peng Chu, Xiao Bian, Shaopeng Liu, and Haibin Ling. Feature space augmentation for long-tailed data. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision – ECCV 2020*, pp. 694–710, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58526-6.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. All that's 'human' is not gold: Evaluating human evaluation of generated text. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 7282–7296, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.565. URL https://aclanthology.org/2021.acl-long.565.

Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A. Smith, and Yejin Choi. Scarecrow: A framework for scrutinizing machine text, 2021.

Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A. Smith, and Yejin Choi. Is GPT-3 text indistinguishable from human text? scarecrow: A framework for scrutinizing machine

text. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7250–7274, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.501. URL https://aclanthology.org/2022.acl-long.501.

Mor Geva, Yoav Goldberg, and Jonathan Berant. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 1161–1166, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1107. URL https://aclanthology.org/D19-1107.

Morgan Gray, Jaromir Savelka, Wesley Oliver, and Kevin Ashley. Can gpt alleviate the burden of annotation? In *Legal Knowledge and Information Systems*, pp. 157–166. IOS Press, 2023.

Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. LongT5: Efficient text-to-text transformer for long sequences. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 724–736, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.55. URL https://aclanthology.org/2022.findings-naacl.55.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 107–112, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2017. URL https://aclanthology.org/N18-2017.

Sophie Henning, William Beluch, Alexander Fraser, and Annemarie Friedrich. A survey of methods for addressing class imbalance in deep-learning based natural language processing. In Andreas Vlachos and Isabelle Augenstein (eds.), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 523–540, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.38. URL https://aclanthology.org/2023.eacl-main.38.

Yushi Hu, Chia-Hsuan Lee, Tianbao Xie, Tao Yu, Noah A. Smith, and Mari Ostendorf. In-context learning for few-shot dialogue state tracking. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 2627–2643, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-emnlp.193.

J. F. Kelley. An iterative design methodology for user-friendly natural language office information applications. *ACM Trans. Inf. Syst.*, 2(1):26–41, jan 1984. ISSN 1046-8188. doi: 10.1145/357417.357420. URL https://doi.org/10.1145/357417.357420.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, et al. Soda: Million-scale dialogue distillation with social commonsense contextualization. *arXiv preprint arXiv:2212.10465*, 2022.

Sophia Lam, Charles Chen, Kristi Kim, George Wilson, J Holt Crews, and Matthew S Gerber. Optimizing customer-agent interactions with natural language processing and machine learning. In *2019 Systems and Information Engineering Design Symposium (SIEDS)*, pp. 1–6. IEEE, 2019.

Chia-Hsuan Lee, Hao Cheng, and Mari Ostendorf. Dialogue state tracking with a language model using schema-driven prompting. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4937–4949, Online and Punta Cana, Dominican

Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.404. URL https://aclanthology.org/2021.emnlp-main.404.

Haonan Li, Martin Tomko, Maria Vasardani, and Timothy Baldwin. MultiSpanQA: A dataset for multi-span question answering. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1250–1260, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.90. URL https://aclanthology.org/2022.naacl-main.90.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. Synthetic data generation with large language models for text classification: Potential and limitations. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10443–10461, Singapore, December 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.emnlp-main.647.

Joseph CR Licklider. Man-computer symbiosis. *IRE transactions on human factors in electronics*, (1):4–11, 1960.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. WANLI: Worker and AI collaboration for natural language inference dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 6826–6847, Abu Dhabi, United Arab Emirates, December 2022a. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-emnlp.508.

Fan Liu, Zhiyong Cheng, Huilin Chen, Yinwei Wei, Liqiang Nie, and Mohan Kankanhalli. Privacy-preserving synthetic data generation for recommendation systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, pp. 1379–1389, New York, NY, USA, 2022b. Association for Computing Machinery. ISBN 9781450387323. doi: 10.1145/3477495.3532044. URL https://doi.org/10.1145/3477495.3532044.

Bo-Ru Lu, Yushi Hu, Hao Cheng, Noah A. Smith, and Mari Ostendorf. Unsupervised learning of hierarchical conversation structure. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 5657–5670, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.415. URL https://aclanthology.org/2022.findings-emnlp.415.

Bo-Ru Lu, Nikita Haduong, Chien-Yu Lin, Hao Cheng, Noah A. Smith, and Mari Ostendorf. Encode once and decode in parallel: Efficient transformer decoding, 2024.

OpenAI. Gpt-4 technical report, 2024.

Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–18, 2022.

Noseong Park, Mahmoud Mohammadi, Kshitij Gorde, Sushil Jajodia, Hongkyu Park, and Youngmin Kim. Data synthesis based on generative adversarial networks. *Proc. VLDB Endow.*, 11(10):1071–1083, jun 2018. ISSN 2150-8097. doi: 10.14778/3231751.3231757. URL https://doi.org/10.14778/3231751.3231757.

Piotr Pęzik, Gosia Krawentek, Sylwia Karasińska, Paweł Wilk, Paulina Rybińska, Anna Cichosz, Angelika Peljak-Łapińska, Mikołaj Deckert, and Michał Adamczyk. DiaBiz – an annotated corpus of Polish call center dialogs. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 723–726, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.76.

Kun Qian, Ahmad Beirami, Zhouhan Lin, Ankita De, Alborz Geramifard, Zhou Yu, and Chinnadhurai Sankar. Annotation inconsistency and entity bias in MultiWOZ. In Haizhou Li, Gina-Anne Levow, Zhou Yu, Chitralekha Gupta, Berrak Sisman, Siqi Cai, David Vandyke, Nina Dethlefs, Yan Wu, and Junyi Jessy Li (eds.), *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 326–337, Singapore and Online, July 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.sigdial-1.35. URL https://aclanthology.org/2021.sigdial-1.35.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67, 2020.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL https://aclanthology.org/D16-1264.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, number 05, pp. 8689–8696, 2020.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410.

Ashish Sharma, Inna Wanyin Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. Human–ai collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence*, 5:46–57, 2022. URL https://api.semanticscholar.org/CorpusID:247778407.

Felix Stahlberg and Shankar Kumar. Synthetic data generation for grammatical error correction with tagged corruption models. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 37–47, Online, April 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.bea-1.4.

Ramya Tekumalla and Juan M. Banda. Leveraging large language models and weak supervision for social media data annotation: An evaluation using covid-19 self-reported vaccination tweets. In Hirohiko Mori, Yumi Asahi, Adela Coman, Simona Vasilache, and Matthias Rauterberg (eds.), *HCI International 2023 – Late Breaking Papers*, pp. 356–366, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-48044-7.

Amirsina Torfi, Edward A. Fox, and Chandan K. Reddy. Differentially private synthetic medical data generation using convolutional gans. *Information Sciences*, 586:485–500, 2022. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2021.12.018. URL https://www.sciencedirect.com/science/article/pii/S0020025521012391.

Adaku Uchendu, Jooyoung Lee, Hua Shen, Thai Le, Dongwon Lee, et al. Does human collaboration enhance the accuracy of identifying llm-generated deepfake texts? In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 11, pp. 163–174, 2023.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13484–13508, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.754. URL https://aclanthology.org/2023.acl-long.754.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Bangzhou Xin, Wei Yang, Yangyang Geng, Sheng Chen, Shaowei Wang, and Liusheng Huang. Private fl-gan: Differential privacy synthetic data generation based on federated learning. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2927–2931, 2020. doi: 10.1109/ICASSP40776.2020.9054559.

Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Zhipeng Zhao, Kun Zhou, Xiaolei Wang, Wayne Xin Zhao, Fan Pan, Zhao Cao, and Ji-Rong Wen. Alleviating the long-tail problem in conversational recommender systems. In *Proceedings of the 17th ACM Conference on Recommender Systems*, RecSys '23, pp. 374–385, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400702419. doi: 10.1145/3604915.3608812. URL https://doi.org/10.1145/3604915.3608812.

# A Evaluation metrics

As described in subsection 3.3, we denote the equations of turn update scores for four diagnostic score types $\text{TYPE} \in \{\text{TLB}, \text{R}, \text{RS}, \text{SV}\}$. All of which are based on averaging over all $N$ dialogues in the test set formulated as

$$\frac{1}{N}\sum_{n}\frac{1}{|\mathcal{T}_n|}\sum_{t\in\mathcal{T}_n} m_{\text{TYPE}}(n,t)$$

where $\mathcal{T}_n = \{t : \mathcal{R}_{nt} \neq \emptyset\}$.

Specifically, the equations of the specific scores ($m_{\text{TYPE}}$) are as follows,

$$\begin{aligned}
m_{\text{TLB}}(n,t) &= \frac{1}{|\mathcal{R}_{nt}|}\sum_{r\in\mathcal{R}_{nt}} m(\hat{\mathbf{S}}_{nrt}, \mathbf{S}^*_{nrt}) \\
m_{\text{R}}(n,t) &= m(\hat{\mathcal{R}}_{nt}, \mathcal{R}^*_{nt}) \\
m_{\text{RS}}(n,t) &= \frac{1}{|\mathcal{R}_{nt}|}\sum_{r\in\mathcal{R}_{nt}} m(\hat{\mathcal{S}}_{nrt}, \mathcal{S}^*_{nrt}) \\
m_{\text{SV}}(n,t) &= m\left(\bigcup_{r\in\mathcal{R}_{nt}} \hat{\mathbf{S}}_{nrt}, \bigcup_{r\in\mathcal{R}_{nt}} \mathbf{S}^*_{nrt}\right)
\end{aligned}$$

where $\mathcal{S}_{nrt}$ is the set of slot labels associated with referent $r$ in turn $t$ of the $n$-th dialogue.

# B Data statistics

| | AIC | DIALGEN-AIC | AUTOGEN-AIC |
|---|---|---|---|
| # dialogue | 34 | 235 | 195 |
| # turns / dialogue | $197 \pm 98$ | $46 \pm 8$ | $31 \pm 7$ |
| # tokens / dialogue | $4195 \pm 2404$ | $1128 \pm 230$ | $947 \pm 214$ |
| # user tokens / turn | $18 \pm 27$ | $22 \pm 17$ | $27 \pm 17$ |
| # agent tokens / turn | $25 \pm 31$ | $27 \pm 14$ | $35 \pm 16$ |
| # referent-slot pair | 1622 | 8844 | – |
| # unique referent-slot | 109 | 152 | – |
| # referent-slot pair / dialogue | $48 \pm 24$ | $38 \pm 8$ | – |
| % dialogue w/ updates | 50.0% | 14.5% | – |
| % dialogue w/ multiple values | 50.0% | 19.1% | – |

Table 5: Statistics are calculated on the full dataset. Tokens are calculated with Huggingface T5 tokenizer.

Table 5 shows the detailed statistics of three datasets: AIC, DIALGEN-AIC, AUTOGEN-AIC. For AUTOGEN, we only generate and annotate the training set with LMs; thus the statistics based on human annotations are not available.

# C Training and generation details

## C.1 Finetuning details

All experiments are done with T5-base or Long-T5-base with Huggingface implementation (Wolf et al., 2020). Training time for full DIALGEN-AIC and AIC setting is averaged 3 hours on 2 NVIDIA V100 GPUs. For the experiments on only DIALGEN-AIC, we use 2 NVIDIA A40 GPUs. The total number of GPU training hours is 110 hours.

| Hyperparameter | SBERT | T5 or T5-SC | Long-T5 |
|---|---|---|---|
| Training batch size | 32 | 16 | 16 |
| Learning rate | $2 \times 10^{-5}$ | $5 \times 10^{-4}$ | $5 \times 10^{-4}$ |
| Max generation length | – | 256 | 256 |
| Max input length | 512 | 512 | 2592 |
| Pretrained Checkpoints | `all-mpnet-base-v2` | `t5-base` | `long-t5-tglobal-base` |

Table 6: Hyperparameters for training in-context learning retriever SBERT and finetuned models T5, T5-SC and Long-T5. The other parameters are default values in Huggingface trainer.

### C.2 Hyperparameters of OpenAI API calls

| Hyperparameter | DIALGEN | AUTOGEN | LM annotators | IC-DST |
|---|---|---|---|---|
| Language Model | ChatGPT | ChatGPT | ChatGPT or GPT-4 | ChatGPT |
| Temperature | 0.85 - 0.9 | 0.0 | 0.9 | 0.0 |
| Max tokens | 512 | 256 | 2048 | 512 |
| Stop strings | ["<\div>"] | – | – | ["–", "\n", ";", "#"] |
| Presence penalty | 0.2 | 0.0 | 0.2 | 0 |
| Frequency penalty | 0.2 | 0.0 | 0.2 | 0 |

Table 7: Hyperparameters for API calls. We use `gpt-3.5-turbo-0301` and `gpt-4-0613` for ChatGPT and GPT-4, respectively. For AUTOGEN, the stop token is the default value of the API call. We parse the output of the response from ChatGPT and truncate tokens after the string `[End of conversation]`.

## D Prompts

We show the prompts used in DIALGEN and AUTOGEN for generating DIALGEN-AIC, AUTOGEN-AIC, IC-DST, T5, T5-SC and Long-T5 in the following subsections.

| Instruction | Count |
|---|---|
| Have CALLER describe more car accident details with complex reasoning that involves two cars' motion. | 23 |
| Have CALLER's response be less specific. have AGENT asks for more details. | 18 |
| Split AGENT's questions into multiple turns | 18 |
| Have CALLER's response be less specific. have AGENT asks for more details. have AGENT asks a question for car accident details. | 15 |
| Have AGENT ask for permission to record the call. | 15 |
| Ask for email address and home address | 14 |
| Have CALLER ask AGENT questions about her insurance coverages in multiple turns | 13 |
| Have AGENT ask CALLER more questions about the accident details | 12 |
| Have CALLER misremember the details. AGENT double check with CALLER. | 12 |
| Explain coverages | 12 |
| Have CALLER corrects wrong information. have AGENT asks for clarification. | 12 |
| Break this conversation down into multiple turns of dialogue | 11 |
| Have AGENT ask for contact information | 10 |
| Break these turns down into multiple turns of back and forth dialogue | 10 |
| AGENT needs to split up her questions. | 10 |

Table 8: Instructions with a frequency of 10 or more times used by humans to regenerate a subdialogue.

| **Personality** | **Description** |
|---|---|
| Aggressive | Feeling angry and confrontational about the accident, may place blame on others or use aggressive language. |
| Analytical | Focused on the details and logistics of the claim process, may ask for precise information and explanations. |
| Confused | Unsure about what happened during the accident or what to do next, may ask a lot of questions. |
| Cooperative | Willing to work with the insurance company and other parties involved in resolving the claim. |
| Defensive | Feeling the need to justify their actions or place blame on others, may be unwilling to take responsibility for the accident. |
| Emotional | Experiencing strong emotions related to the accident, may be crying or struggling to maintain composure during the call. |
| Evasive | Hesitant to provide information or answer questions about the accident, may be trying to conceal something. |
| Impatient | Feeling frustrated with the claim process or the speed at which it is progressing, may express irritation or urgency in their language. |
| Reassuring | Trying to maintain a positive and optimistic outlook during the call, may express gratitude for the assistance being provided. |
| Upset | Feeling distressed or frustrated due to the accident and its consequences. |

Table 9: The list of the predefined callers' personalities.

### D.1 DIALGEN prompts

**Prompts for instructing a language model to generate a desired subdialogue.** Table 8 shows the instructions used by humans when regenerating a subdialogue. The top four instructions instruct the LM to produce more content, which naturally results in lengthier dialogues.

**Prompts for providing callers' personality.** Table 9 shows the list of predefined callers' personality. We randomly select one personality when creating the dialogue scenario.

**Prompts for providing scenarios and dialogue content to generate a subdialogue.** We present an example prompt used to generate the first subdialogue when using DIALGEN-AIC for auto insurance claim calls. It includes a task description, entity-slot-value triplets, an accident story, caller's and agent's personalities, and an initial exchange. Similar to Park et al. (2022), we use HTML tags to denote different dialogue elements, i.e., <p> for turns and  for the subdialogue.

```
<short_summary>
story
Bob Parkhurst had a busy day at work, and all he wanted to do was to go grocery shopping. As he backed out of her
 parking spot in the Office Depot parking lot, he failed to notice the gray MAZDA B-Series Extended Cab driven by
 Spencer Tullar as he turned into the same aisle from the opposite direction.
Spencer, who was on his way to run some errands, had been driving down the parking lot in extremely slow speed
when suddenly he saw Bob's yellow car backing out of his spot. He didn't think much of it and was about to just
drive behind her when, at the last minute, he noticed that Bob seemed to be backing out without looking around.
Spencer slammed on his brakes, but it was too late. The front right of his truck smashed hard into the back
passenger side of Bob's car.
The impact of the collision caused Bob's car to spin around and come to a stop. He immediately felt a sharp pain
in her neck and knew that something was wrong. As he tried to get out of the car, he realized that he couldn't
move his neck without experiencing excruciating pain.
Spencer got out of his truck and approached Bob's car, he asked if Bob was okay. Bob told him that he was hurt
and needed medical attention. Spencer called 911 immediately while also trying his best to comfort Bob until help
 arrived.
When emergency services arrived shortly after, they found Bob slumped over in her seat, clutching his neck in
agony. The responders helped her out of the car and placed a neck brace around him so he wouldn't move his head
while they examined her injuries. They then transported him by ambulance to the hospital for further medical
attention.
Meanwhile, police were already on their way. Upon arrival at the scene, they took statements from both drivers as
 well as any witnesses who may have seen what happened. Unfortunately, no one at the time had a clear view of the
 incident, but both drivers agreed that they didn't see each other before the collision.
Since both cars were still in the parking lot when the accident happened, there was no need to redirect traffic.
However, the officers still had to direct people away from the incident site to prevent any further accidents.
They also checked Spencer's license and found that it was valid.
The investigation into what caused the accident was inconclusive. Neither driver was certain about who was at
fault, as they both believed the other driver failed to observe their movements. Since no one appeared to be at
fault, no tickets or
--------
entity-slot-value triplets
Accident details: (accident location, office depot parking lot), (damage part, unsure), num of passengers,
witnesses, date of accident, time of accident, subjective fault, airbag deployed.
Evidences of the car accident: police report, (pictures, no picture), police report number, police department
name, tickets citations.
Traffic condition: weather visibility, (obstructions to view, no).
Caller's driver action: car motion, speed, traffic controls obeyed, turn signal, (horn, no).
Caller's car information: (make/model, dodge stratus), make year, color, car mileage.
Caller's injury details: body part injured, injury type, medical treatment.
--------
task description
Have role play car accident claim call. One person is an agent Alice from a car insurance company and the other
is the caller Bob who wants to file a claim.
At beginning of the call, have Alice ask for Bob's permission to record the call and proceeds with the
conversation.
Within some <p> </p>, have simulate poor phone connection. Have Alice and Bob can not hear each other and need to
 repeat what they said.
Have Alice verify Bob personal information to access account information at the beginning of the call.
Have Bob describe the car accident by using story and tuples above to describe the accident.
Have Alice confirm new information with Bob during the call to ensure consistency.
Have Alice and Bob engage in small talk with each other.
Have Alice explain the insurance coverages to Bob.
--------
personality
Bob is impatient, feeling frustrated with the claim process or the speed at which it is progressing, may express
irritation or urgency in their language.
Alice is conversational, personable, patient, empathetic, sympathetic and professional.
--------
```

```
instructions
Use the story, information, and personality to create a role play script and follow the task description.
</short_summary>
<div>
<p class="Alice" title="Auto Accident"> Thank you for calling! This is Alice. How may I help you today? </p>
<p class="Bob" title="Auto Accident"> Hello. This is Alice. I am calling for a car accident. </p>
</div>
Have Alice ask a question for car accident details.
<div>
```

### D.2 AUTOGEN prompts

We use two separate prompts for generating an AUTOGEN-AIC dialogue and doing automatic annotation.

**Prompts for automatic dialogue generation.** The AUTOGEN-AIC prompt requires two things: a scenario and a demonstration dialogue. The demonstration dialogue provides the language model (LM) the format and style of the desired dialogue, and the scenario describes the essential information the generated dialogue should cover. We select a dialogue from the DIALGEN-AIC training set as the demonstration dialogue and share it across all generated dialogues. To compare DIALGEN and AUTOGEN, we use 195 scenarios from DIALGEN training dialogues and apply them to AUTOGEN to automatically generate dialogues. The LM generates a full dialogue following the format in the demonstration dialogue and the information in the scenario. We cut the generated output after the phrase `[end of the dialogue]`. The output before the cut is the final generated dialogue. The following python string is used as the template.

```
# Scenario\n{scenario}
------------------
# Demo dialogue\n{demo_dialogue}
[end of the dialogue]
------------------
Generate a new dialogue between an agent and a user for a car accident claim based on the information mention in
the scenario.
Follow the format and style of the demo dialogue.
Agent:
```

**Prompts for automatic annotation.** We use ChatGPT or GPT-4 to automatically annotate a pair of system user turns. `[SYSTEM TURN]` and `[USER TURN]` are the placeholders for user and system turns.

```
Possible Entities: Global, Caller, Other Driver, Caller's Passenger, Other Driver's Passenger, Witness.
Entity description:
Global: it is for slot values that are not associated with any entity.
Caller: the slot values that are associated with the caller.
Other Driver: the slot values that are associated with the other driver.
Caller's Passenger: the slot values that are associated with the caller's passenger.
Other Driver's Passenger: the slot values that are associated with the other driver's passenger.
Witness: the slot values that are associated with the witness.

---------------
Schema table:
domain-slot   possible values
CarInfo-Make/Model    non-categorical values
CarInfo-Make Year     non-categorical values
CarInfo-Color  non-categorical values
CarInfo-Car Mileage   non-categorical values
CarInfo-Rideshare (Uber/Lyft) yes, no, unsure
Adjuster-Explain Coverages   non-categorical values
Adjuster-Permission to Record yes, no
Adjuster-Set up Inspection   photo claim, field assignment
Adjuster-Set up Rental yes, no
TrafficEnvironment-Weather Visibility clear, cloudy, rainy, snowy, foggy, windy, other, unsure
TrafficEnvironment-Obstructions to View yes, no, unsure
TrafficEnvironment-Road Condition   dry, wet, slippery, debris, potholes, straight, curved, tunnel, steep incline,
 flat, other, unsure
TrafficEnvironment-Traffic Signal   stop sign, yield sign, green light, yellow light, red light, other, unsure,
no signal or sign
TrafficEnvironment-Description of Lanes normal, turn lane, shoulder, other, unsure
TrafficEnvironment-Num of Lanes 1, 2, 3, 4+, unsure
TrafficEnvironment-Traffic Condition heavy, moderate, light, other, unsure
```

```
TrafficEnvironment-Speed Limit non-categorical values
TrafficEnvironment-Traffic Flow one-way, two-way, other, unsure
TrafficEnvironment-Parking Lot Type angled, straight, other, unsure
AccidentDetails-Damage Part  front, right, back, left, front right, front left, back left, back right, other,
unsure
AccidentDetails-Accident Location   parking lot, driveway, highway, roadway, intersection, other
AccidentDetails-Num of Passengers   0, 1, 2+, unsure
AccidentDetails-Witnesses    yes, no, unsure
AccidentDetails-Num of Involved Cars 1, 2, 3, 4+, unsure
AccidentDetails-Children Involved   yes, no, unsure
AccidentDetails-Airbag Deployed yes, no, unsure
AccidentDetails-Towed yes, no, unsure
AccidentDetails-Pedestrians Involved yes, no, unsure
AccidentDetails-Date of Accident    non-categorical values
AccidentDetails-Time of Accident    non-categorical values
AccidentDetails-Subjective Fault    caller, other driver
ContactInfo-First Name non-categorical values
ContactInfo-Last Name non-categorical values
ContactInfo-Home Address     non-categorical values
ContactInfo-Phone Number     non-categorical values
ContactInfo-Email Address    non-categorical values
ContactInfo-Policy Number    non-categorical values
ContactInfo-Date of Birth    non-categorical values
DriverActions-Car Motion     traveling forward, backing, turning, changing lanes, stopped, other, unsure
DriverActions-Speed  non-categorical values
DriverActions-Distractions   cellphone, animals, smoking, passengers, traffic, eating, not paying attention,
other, unsure, no distraction
DriverActions-Brake   yes, no, unsure
DriverActions-Horn    yes, no, unsure
DriverActions-Turn Signal    yes, no, unsure
DriverActions-Traffic Controls Obeyed yes, no, unsure
Trip-Destination of Trip     non-categorical values
Trip-Purpose of Trip  non-categorical values
Trip-Origin of Trip   non-categorical values
InjuryDetails-Ambulance yes, no, unsure
InjuryDetails-Body Part Injured head, neck, shoulder, chest, abdomen, back, limb, other
InjuryDetails-Injury Type    bruise, broken fracture, cut scratch, bleeding, strain sprain, sore, other, no
injury
InjuryDetails-Medical Treatment MRI, surgery, CAT scan, hospitalization, ER, x-ray, other
Evidences-Police Report yes, no, unsure
Evidences-Police Department Name    non-categorical values
Evidences-Pictures    at scene, after accident, no picture, unsure
Evidences-Tickets Citations  caller party cited, other party cited, no party cited, multiple parties cited,
unsure, no ticket
Evidences-Police Report Number non-categorical values
Evidences-Skid Marks  yes, no, unsure
---------------
Demonstration
Example 1
system turn: Great, thank you for providing that information, Patrick. Can you please tell me about the accident?
 When and where did it happen?
user turn: The accident occurred at 5th Avenue and Main Street yesterday around 4 pm. I was driving on Main
Street when a Ford E350 Super Duty Cargo backed up from a turn lane and hit my car's right side.
tlb label:
{
  "AccidentDetails-Accident Location": [
    "Global || Intersection"
  ],
  "AccidentDetails-Date of Accident": [
    "Global || yesterday"
  ],
  "AccidentDetails-Time of Accident": [
    "Global || 4 pm."
  ],
  "AccidentDetails-Damage Part": [
    "Caller || Right"
  ],
  "CarInfo-Make/Model": [
    "Other Driver || Ford E350 Super Duty Cargo"
  ],
  "DriverActions-Car Motion": [
    "Caller || Traveling Forward",
    "Other Driver || Backing"
  ],
  "TrafficEnvironment-Description of Lanes": [
    "Global || Turn Lane"
  ]
}


Example 2
```

```
system turn: Absolutely, Patrick. You have collision insurance, which will cover the damages to your car from the
 accident. You also have uninsured/underinsured motorist coverage in case the other driver doesn't have enough
insurance to cover the damages. Your MedPay coverage will pay for any medical expenses resulting from the
accident. And finally, you have comprehensive insurance, which covers damages not related to a collision, such as
 theft or natural disasters.
user turn: That sounds good, Debra. Can you tell me more about MedPay?
tlb label:
{
  "Adjuster-Explain Coverages": [
    "Global || You have collision insurance, which will cover the damages to your car from the accident. You also
     have uninsured/underinsured motorist coverage in case the other driver doesn't have enough insurance to cover
     the damages. Your MedPay coverage will pay for any medical expenses resulting from the accident. And finally,
     you have comprehensive insurance, which covers damages not related to a collision, such as theft or natural
    disasters."
  ]
}


---------------
Follow the demonstration to label the slot values for the following system and user turns. Only return extracted
slot values. Use double quotes for keys and values in the returned dictionary. Return all labels in a single
dictionary. Make sure the to return a correct format of the dictionary.

system turn: [SYSTEM TURN]
user turn: [USER TURN]
tlb label:
```

### D.3 IC-DST prompt and output

Due to the input length limit, we extract the TLB at turn $t$ and accumulate TLBs as CB. Thus, `[context]` is empty.

```
CREATE TABLE AccidentDetails(
    'Damage Part' TEXT CHECK ('Damage Part' IN 'Front', 'Right', 'Back', 'Left', 'Front Right', 'Front Left', '
    Back Left', 'Back Right', 'Other', 'Unsure'),
    'Accident Location' TEXT CHECK ('Accident Location' IN 'Parking Lot', 'Driveway', 'Highway', 'Roadway', '
    Intersection', 'Other'),
    'Num of Passengers' TEXT CHECK ('Num of Passengers' IN '0', '1', '2+', 'Unsure'),
    'Witnesses' TEXT CHECK ('Witnesses' IN 'Yes', 'No', 'Unsure'),
    'Num of Involved Cars' TEXT CHECK ('Num of Involved Cars' IN '1', '2', '3', '4+', 'Unsure'),
    'Children Involved' TEXT CHECK ('Children Involved' IN 'Yes', 'No', 'Unsure'),
    'Airbag Deployed' TEXT CHECK ('Airbag Deployed' IN 'Yes', 'No', 'Unsure'),
    'Towed' TEXT CHECK ('Towed' IN 'Yes', 'No', 'Unsure'),
    'Pedestrians Involved' TEXT CHECK ('Pedestrians Involved' IN 'Yes', 'No', 'Unsure'),
    'Date of Accident' TEXT,
    'Time of Accident' TEXT,
    'Subjective Fault' TEXT CHECK ('Subjective Fault' IN 'Caller', 'Other Driver'),
)

CREATE TABLE Adjuster(
    'Explain Coverages' TEXT,
    'Permission to Record' TEXT CHECK ('Permission to Record' IN 'Yes', 'No'),
    'Set up Inspection' TEXT CHECK ('Set up Inspection' IN 'Quick Photo Claim', 'Field Assignment'),
    'Set up Rental' TEXT CHECK ('Set up Rental' IN 'Yes', 'No'),
)

CREATE TABLE CarInfo(
    'Make/Model' TEXT,
    'Make Year' TEXT,
    'Color' TEXT,
    'Car Mileage' TEXT,
    'Rideshare (Uber/Lyft)' TEXT CHECK ('Rideshare (Uber/Lyft)' IN 'Yes', 'No', 'Unsure'),
)

CREATE TABLE ContactInfo(
    'First Name' TEXT,
    'Last Name' TEXT,
    'Home Address' TEXT,
    'Phone Number' TEXT,
    'Email Address' TEXT,
    'Policy Number' TEXT,
    'Date of Birth' TEXT,
)

CREATE TABLE DriverActions(
    'Car Motion' TEXT CHECK ('Car Motion' IN 'Traveling Forward', 'Backing', 'Turning', 'Changing Lanes', '
    Stopped', 'Other', 'Unsure'),
    'Speed' TEXT,
```

```
    'Distractions' TEXT CHECK ('Distractions' IN 'Cellphone', 'Animals', 'Smoking', 'Passengers', 'Traffic', '
    Eating', 'Not Paying Attention', 'Other', 'Unsure', 'No Distraction'),
    'Brake' TEXT CHECK ('Brake' IN 'Yes', 'No', 'Unsure'),
    'Horn' TEXT CHECK ('Horn' IN 'Yes', 'No', 'Unsure'),
    'Turn Signal' TEXT CHECK ('Turn Signal' IN 'Yes', 'No', 'Unsure'),
    'Traffic Controls Obeyed' TEXT CHECK ('Traffic Controls Obeyed' IN 'Yes', 'No', 'Unsure'),
)

CREATE TABLE Evidences(
    'Police Report' TEXT CHECK ('Police Report' IN 'Yes', 'No', 'Unsure'),
    'Police Department Name' TEXT,
    'Pictures' TEXT CHECK ('Pictures' IN 'At Scene', 'After Accident', 'No Picture', 'Unsure'),
    'Tickets Citations' TEXT CHECK ('Tickets Citations' IN 'Caller Party Cited', 'Other Party Cited', 'No Party
    Cited', 'Multiple Parties Cited', 'Unsure', 'No Ticket'),
    'Police Report Number' TEXT,
    'Skid Marks' TEXT CHECK ('Skid Marks' IN 'Yes', 'No', 'Unsure'),
)

CREATE TABLE InjuryDetails(
    'Ambulance' TEXT CHECK ('Ambulance' IN 'Yes', 'No', 'Unsure'),
    'Body Part Injured' TEXT CHECK ('Body Part Injured' IN 'Head', 'Neck', 'Shoulder', 'Chest', 'Abdomen', 'Back',
     'Limb', 'Other'),
    'Injury Type' TEXT CHECK ('Injury Type' IN 'Bruise', 'Broken Fracture', 'Cut Scratch', 'Bleeding', 'Strain
    Sprain', 'Sore', 'Other', 'No Injury'),
    'Medical Treatment' TEXT CHECK ('Medical Treatment' IN 'MRI', 'Surgery', 'Cat Scan', 'Hospitalization', 'ER',
     'X-Ray', 'Other'),
)

CREATE TABLE TrafficEnvironment(
    'Weather Visibility' TEXT CHECK ('Weather Visibility' IN 'Clear', 'Cloudy', 'Rainy', 'Snowy', 'Foggy', 'Windy
    ', 'Other', 'Unsure'),
    'Obstructions to View' TEXT CHECK ('Obstructions to View' IN 'Yes', 'No', 'Unsure'),
    'Road Condition' TEXT CHECK ('Road Condition' IN 'Dry', 'Wet', 'Slippery', 'Debris', 'Potholes', 'Straight',
    'Curved', 'Tunnel', 'Steep Incline', 'Flat', 'Other', 'Unsure'),
    'Traffic Signal' TEXT CHECK ('Traffic Signal' IN 'Stop Sign', 'Yield Sign', 'Green Light', 'Yellow Light', '
    Red Light', 'Other', 'Unsure', 'No Signal Or Sign'),
    'Description of Lanes' TEXT CHECK ('Description of Lanes' IN 'Normal', 'Turn Lane', 'Shoulder', 'Other', '
    Unsure'),
    'Num of Lanes' TEXT CHECK ('Num of Lanes' IN '1', '2', '3', '4+', 'Unsure'),
    'Traffic Condition' TEXT CHECK ('Traffic Condition' IN 'Heavy', 'Moderate', 'Light', 'Other', 'Unsure'),
    'Speed Limit' TEXT,
    'Traffic Flow' TEXT CHECK ('Traffic Flow' IN 'One-Way', 'Two-Way', 'Other', 'Unsure'),
    'Parking Lot Type' TEXT CHECK ('Parking Lot Type' IN 'Angled', 'Straight', 'Other', 'Unsure'),
)

CREATE TABLE Trip(
    'Destination of Trip' TEXT,
    'Purpose of Trip' TEXT,
    'Origin of Trip' TEXT,
)

-- Using valid SQLite, answer the following multi-turn conversational questions for the tables provided above.

Example #1
[context]
[system] I see. Thank you for letting me know. Can you also provide me with the make, model, and year of your car,
 as well as its color?
Q: [user] Of course. It's a white Lexus sedan, 2018 model.
SQL: SELECT * FROM CarInfo WHERE Caller-Make_Year = 2018 AND Caller-Color = white AND Caller-Make/Model = Lexus
sedan,;


Example #2
[context]
[system] Thank you for sharing that information, Lynne. Can you also provide me with the make and model of your
car?
Q: [user] Yes, it's a white sedan. The make and model is a Toyota Camry. It's a 2018 model, and it had about
40,000 miles on it at the time of the accident
.
SQL: SELECT * FROM CarInfo WHERE Caller-Color = white sedan. AND Caller-Make/Model = Toyota Camry. AND Caller-
Make_Year = 2018 AND Caller-Car_Mileage = 40,
000;


Example #3
[context]
[system] I see. Can you describe your car's make and model? What year was it made? And what color was it?
Q: [user] It's a white sedan, a 2018 Honda Accord.
SQL: SELECT * FROM CarInfo WHERE Caller-Make/Model = sedan, a 2018 Honda Accord. AND Caller-Make_Year = 2018 AND
Caller-Color = white;
```

```
Example #4
[context]
[system] Do you remember the make and model of the other car?
Q: [user] I think it was a black sedan, but I'm not completely sure.
SQL: SELECT * FROM CarInfo WHERE Other_Driver-Make/Model = sedan, AND Other_Driver-Color = black;


Example #5
[context]
[system] Thank you for that information, Joel. Can you please provide me with your car's make and model, year,
color, and approximate mileage?
Q: [user] Sure, my car is a white sedan. It's a 2016 model with approximately 50,000 miles on it.
SQL: SELECT * FROM CarInfo WHERE Caller-Make/Model = sedan. AND Caller-Car_Mileage = approximately 50,000 miles
AND Caller-Color = white AND Caller-Make_Ye
ar = 2016 model;

Example #6
[context]
[system] Thank you for all the details, Richard. Can you please provide me with your car's make and model?
Q: [user] Yes, it's a white sedan, a 2007 make.
SQL: SELECT * FROM
CarInfo WHERE Caller-Color = white sedan AND Caller-Make_Year = 2007
 * FROM CarInfo WHERE Caller-Color = white sedan AND Caller-Make_Year = 2007
 * FROM CarInfo WHERE Caller-Color = white sedan AND Caller-Make_Year = 2007
```

### D.4 Prompt and output for finetuned models

The previous study (Lee et al., 2021) employs independent decoding with natural language prompts for optimal outcomes. However, this approach necessitates the enumeration of all potential combinations of domain-slot pairs during both training and inference. As the ontology grows larger, the computational burden increases linearly. To address this issue, we propose to group slots with the same domain and train the models to predict all active slots with their values and referents simultaneously.

**Long-T5 for CB prediction.** We present a training example for the "ContactInfo" domain with a complete dialogue history at time $t$. The example contains separators `[s]`, `[rv]`, and `[srv]` that label prior information as a slot, referent-value pair, or slot-referent-value triplet, respectively.

```
Input:
[USER] My name is Bob Lee, and my policy number is 123456789. [SYSTEM] Thank you. Could you please provide me
with your name and policy number so I can access your account information? [USER] Yes, that's fine. [SYSTEM] I am
 so sorry that happened. Before we begin, may I please have your permission to record this call for quality and
training purposes? [USER] Hello. This is Bob. I am calling for a car accident. [SYSTEM] Thank you for calling
AllState! This is Alice. How may I help you today? [domain] ContactInfo [possible slots] First Name (the First
Name of the ContactInfo) [s] Last Name (the Last Name of the ContactInfo) [s] Home Address (the Home Address of
the ContactInfo) [s] Phone Number (the Phone Number of the ContactInfo) [s] Email Address (the Email Address of
the ContactInfo) [s] Policy Number (the Policy Number of the ContactInfo) [s] Date of Birth (the Date of Birth of
 the ContactInfo)

Output:
First Name [srv] Bob [rv] Caller [s] Last Name [srv] Lee [rv] Caller [s] Policy Number [srv] 123456789. [rv]
Caller
```

**Long-T5 and T5 models for TLB prediction.** We present a training example for the "ContactInfo" domain with the most recent two turns $(A, U)_t$ at time $t$.

```
Input:
[USER] Hi, my name is Bob Lee. I was recently in a car accident and wanted to file a claim. [SYSTEM] Thank you
for calling! This is Alice. How may I help you today? [domain] ContactInfo [possible slots] First Name (the First
 Name of the ContactInfo) [s] Last Name (the Last Name of the ContactInfo) [s] Home Address (the Home Address of
the ContactInfo) [s] Phone Number (the Phone Number of the ContactInfo) [s] Email Address (the Email Address of
the ContactInfo) [s] Policy Number (the Policy Number of the ContactInfo) [s] Date of Birth (the Date of Birth of
 the ContactInfo)

Output:
First Name [srv] Bob [rv] Caller [s] Last Name [srv] Lee [rv] Caller
```

In the example, the caller (USER) mentions the first and the last name that are under the domain ContactInfo. The model is require to generate the active slots "First Name" and "Last Name" with the corresponding values "Bob" and "Lee", and referent "Caller."

**T5 with State Change (T5-SC).** For T5-SC, the models need to predict entity-slot-value triplets and edit operations associated with the triplets. The final output of a state at time $t$ will be calculated by applying the edit operations on the associated triplets given the previous state at time $t-1$. We consider four edit operations: `[new]`, `[keep]`, `[delete]`, and `[concat]`. We describe the four edit operations in the following paragraph.

If a triplet has not been observed in the previous state, the model is expected to predict `[new]`. Conversely, if the triplet has already been mentioned in the previous state, the model must predict `[keep]`. The `[delete]` operation is employed when a triplet mentioned in the previous state should be removed. If the value of a referent-slot is updated, then the model predicts both `[delete]` for the previous value and `[new]` for the updated value. On the other hand, the `[concat]` operation is used when the value of a triplet needs refinement, such as combining two values, 7 and AM, into a single value of 7 AM.

Due to the input length limit of the T5 model, we use the most recent $k$ turns to create the previous state and omit the slot descriptions in order to cover more entity-slot-value triplets in the previous state. We get the best results when $k = 18$ for DIALGEN-AIC and $k = 20$ for AIC. We present a training example for the "AccidentDetails" domain as follows.

```
Input:
[USER] Oh, sorry about that. You're right, it actually occurred on a Wednesday at 11 am. [SYSTEM] Also, I just
wanted to clarify some information. In our previous conversation, you stated that the accident occurred on a
Monday at 9 am. However, our records show that it actually occurred on a Wednesday at 11 am. Can you confirm
which day and time the accident actually occurred? [state] Damage Part [srv] Front Left [rv] Caller [cv] Right [
rv] Global [s] Accident Location [srv] Highway [rv] Global [s] Num of Passengers [srv] 0 [rv] Global [s]
Witnesses [srv] Yes [rv] Global [s] Date of Accident [srv] this Monday [rv] Global [s] Time of Accident [srv]
9:00 am. [rv] Global [s] Subjective Fault [srv] Caller [rv] Caller [domain] AccidentDetails [possible slots]
Damage Part [s] Accident Location [s] Num of Passengers [s] Witnesses [s] Num of Involved Cars [s] Children
Involved [s] Airbag Deployed [s] Towed [s] Pedestrians Involved [s] Date of Accident [s] Time of Accident [s]
Subjective Fault

Output:
Date of Accident [srv] Wednesday [v] this Monday [vo] [delete] [rv] Global [s] Time of Accident [srv] 11 am. [v]
9:00 am. [vo] [delete] [rv] Global
```

In the example, the agent (SYSTEM) clarifies the date and time with the caller (USER) because the date and time the caller provides are different from the record in the agent's system. The caller admits the provided time and date are wrong. Therefore, the time and date must be updated. The date previously provided "this Monday" needs to be deleted, so we append an operation `[delete]` after the value. Similarly, we append the operation after the time "9:00 am."

# E DIALGEN

## E.1 Data collection cost

The human reviewers were recruited from the university list. They were compensated at a rate of $18.69 per hour following our institution's practices. A dialogue, including reviewing synthesizing and annotation processes, required 45-60 minutes, for a final cost per dialogue of $14-19.

## E.2 IAA

We follow the methodology in SQuAD (Rajpurkar et al., 2016) for calculating IAA. We select 3 trained workers who participated in data generation as our annotators. They annotated 15% of DIALGEN-AIC. The average time to label a dialogue was 18 minutes. For every dialogue, one annotator is randomly assigned as the reference. We calculate max-$F_1$ of every predicted tuple for every turn and average over all turns, then average across all dialogues.

## E.3 AIC ontology

We show the full ontology in Table 10 including domains, slots, and possible values. Possible referents in the AIC ontology: *Global*, *Caller*, *Other Driver*, *Caller's Passenger*, *Other Driver's*

| Domain | Slot | Possible Values |
|---|---|---|
| Adjuster | Explain Coverages | [] |
| Adjuster | Permission to Record | [yes, no] |
| Adjuster | Set up Inspection | [photo claim, field assignment] |
| Adjuster | Set up Rental | [yes, no] |
| ContactInfo | First Name | [] |
| ContactInfo | Last Name | [] |
| ContactInfo | Home Address | [] |
| ContactInfo | Phone Number | [] |
| ContactInfo | Email Address | [] |
| ContactInfo | Policy Number | [] |
| ContactInfo | Date of Birth | [] |
| DriverActions | Car Motion | [traveling forward, backing, turning, changing lanes, stopped, other, unsure] |
| DriverActions | Speed | [] |
| DriverActions | Distractions | [cellphone, animals, smoking, passengers, traffic, eating, not paying attention, other, unsure, no distraction] |
| DriverActions | Brake | [yes, no, unsure] |
| DriverActions | Horn | [yes, no, unsure] |
| DriverActions | Turn Signal | [yes, no, unsure] |
| DriverActions | Traffic Controls Obeyed | [yes, no, unsure] |
| Evidences | Police Report | [yes, no, unsure] |
| Evidences | Police Department Name | [] |
| Evidences | Pictures | [at scene, after accident, no picture, unsure] |
| Evidences | Tickets Citations | [caller party cited, other party cited, no party cited, multiple parties cited, unsure, no ticket] |
| Evidences | Police Report Number | [] |
| Evidences | Skid Marks | [yes, no, unsure] |
| InjuryDetails | Ambulance | [yes, no, unsure] |
| InjuryDetails | Body Part Injured | [head, neck, shoulder, chest, abdomen, back, limb, other] |
| InjuryDetails | Injury Type | [bruise, broken fracture, cut scratch, bleeding, strain sprain, sore, other, no injury] |
| InjuryDetails | Medical Treatment | [MRI, surgery, CAT scan, hospitalization, ER, x-ray, other] |
| AccidentDetails | Damage Part | [front, right, back, left, front right, front left, back left, back right, other, unsure] |
| AccidentDetails | Accident Location | [parking lot, driveway, highway, roadway, intersection, other] |
| AccidentDetails | Num of Passengers | [0, 1, 2+, unsure] |
| AccidentDetails | Witnesses | [yes, no, unsure] |
| AccidentDetails | Num of Involved Cars | [1, 2, 3, 4+, unsure] |
| AccidentDetails | Children Involved | [yes, no, unsure] |
| AccidentDetails | Airbag Deployed | [yes, no, unsure] |
| AccidentDetails | Towed | [yes, no, unsure] |
| AccidentDetails | Pedestrians Involved | [yes, no, unsure] |
| AccidentDetails | Date of Accident | [] |
| AccidentDetails | Time of Accident | [] |
| AccidentDetails | Subjective Fault | [caller, other driver] |
| CarInfo | Make/Model | [] |
| CarInfo | Make Year | [] |
| CarInfo | Color | [] |
| CarInfo | Car Mileage | [] |
| CarInfo | Rideshare (Uber/Lyft) | [yes, no, unsure] |
| Trip | Destination of Trip | [] |
| Trip | Purpose of Trip | [] |
| Trip | Origin of Trip | [] |
| TrafficEnvironment | Weather Visibility | [clear, cloudy, rainy, snowy, foggy, windy, other, unsure] |
| TrafficEnvironment | Obstructions to View | [yes, no, unsure] |
| TrafficEnvironment | Road Condition | [dry, wet, slippery, debris, potholes, straight, curved, tunnel, steep incline, flat, other, unsure] |
| TrafficEnvironment | Traffic Signal | [stop sign, yield sign, green light, yellow light, red light, other, unsure, no signal or sign] |
| TrafficEnvironment | Description of Lanes | [normal, turn lane, shoulder, other, unsure] |
| TrafficEnvironment | Num of Lanes | [1, 2, 3, 4+, unsure] |
| TrafficEnvironment | Traffic Condition | [heavy, moderate, light, other, unsure] |
| TrafficEnvironment | Speed Limit | [] |
| TrafficEnvironment | Traffic Flow | [one-way, two-way, other, unsure] |
| TrafficEnvironment | Parking Lot Type | [angled, straight, other, unsure] |

Table 10: AIC ontology. Empty lists indicate free-form extractive values.

*Passenger*, and *Witness*. All referents could be associated with every domain/slot, although in practice certain information is almost always associated with a particular referent, e.g., Traffic Conditions (heavy, medium, light) always have a *Global* referent.

### E.4 User interface for data collection

We list two main pages of our interface for dialogue generation. They are editing and labeling steps.

First, the editing step (Figure 5) page provides dialogue scenarios (slot value pairs), dialogue history, extracted tuples (annotated entity-slot-value triplets), instruction for regeneration, and current subdialogue for editing. A human reviewer can provide an instruction to guide the LM to generate a desired subdialogue to replace the current subdialogue. If the current subdialogue is satisfied with the reviewer, they can edit turns to fix the minor errors in the subdialogue.

Second, the labeling step page (Figure 6) is an optional page for the DIALGEN framework. This page is designed for the dialogue state tracking task where the human reviewer can annotate and edit the subdialogue in the previous editing step. Note that the labeling step can be fully decoupled from the framework.

The human reviewer will iteratively collaborate with the LM to generate and revise the subdialogues and to annotate the subdialogues until the end of the dialogue.

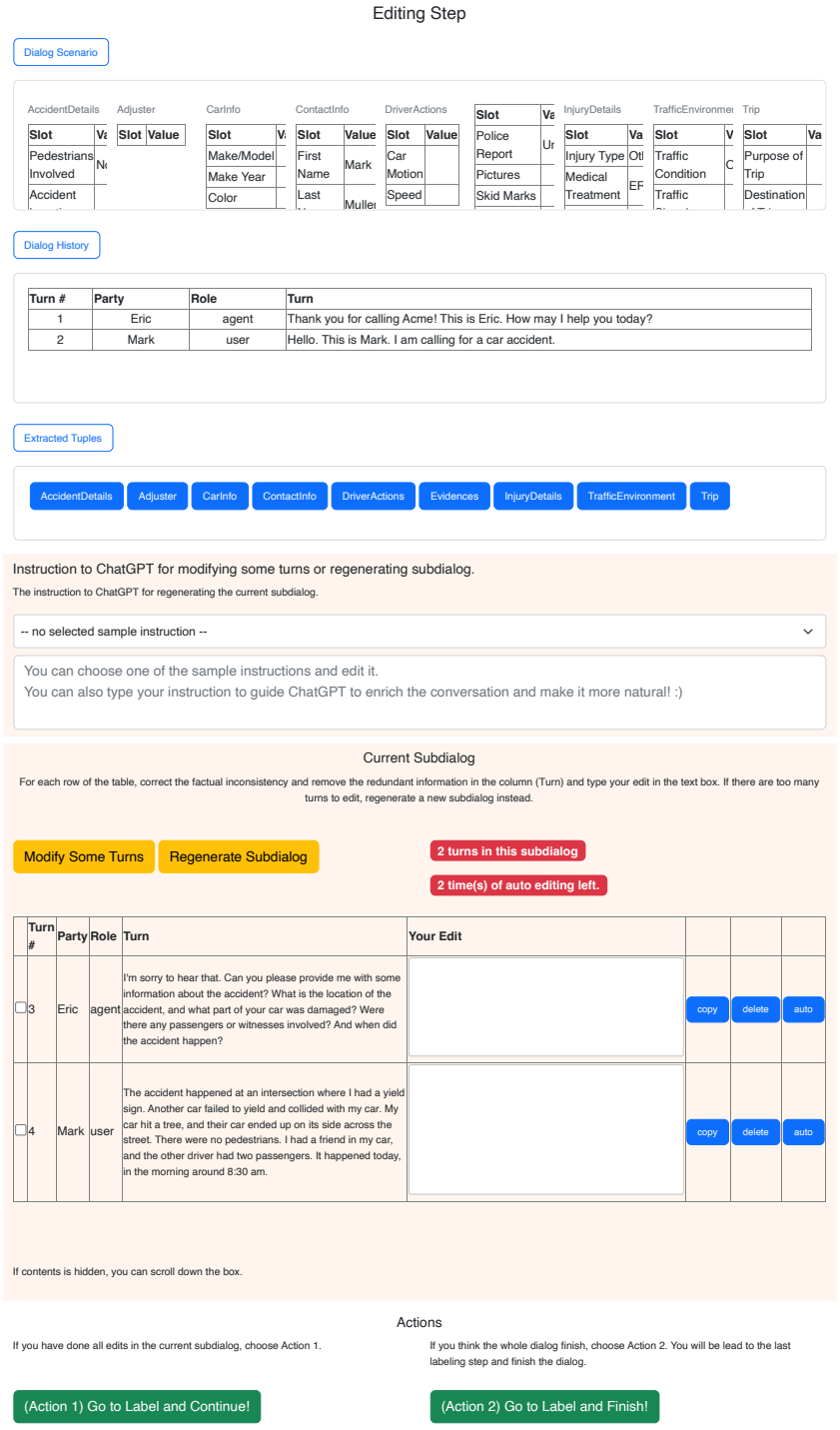


Figure 5: The first step in DIALGEN is to create the subdialogue. A dialogue scenario table is provided to indicate slots expected to appear in the conversation. A human reviewer selects LM-generated text and edit it as needed. They can also ask the LM to regenerate selected turns or the full subdialogue and optionally provide extra instructions to guide the LM's generation process.

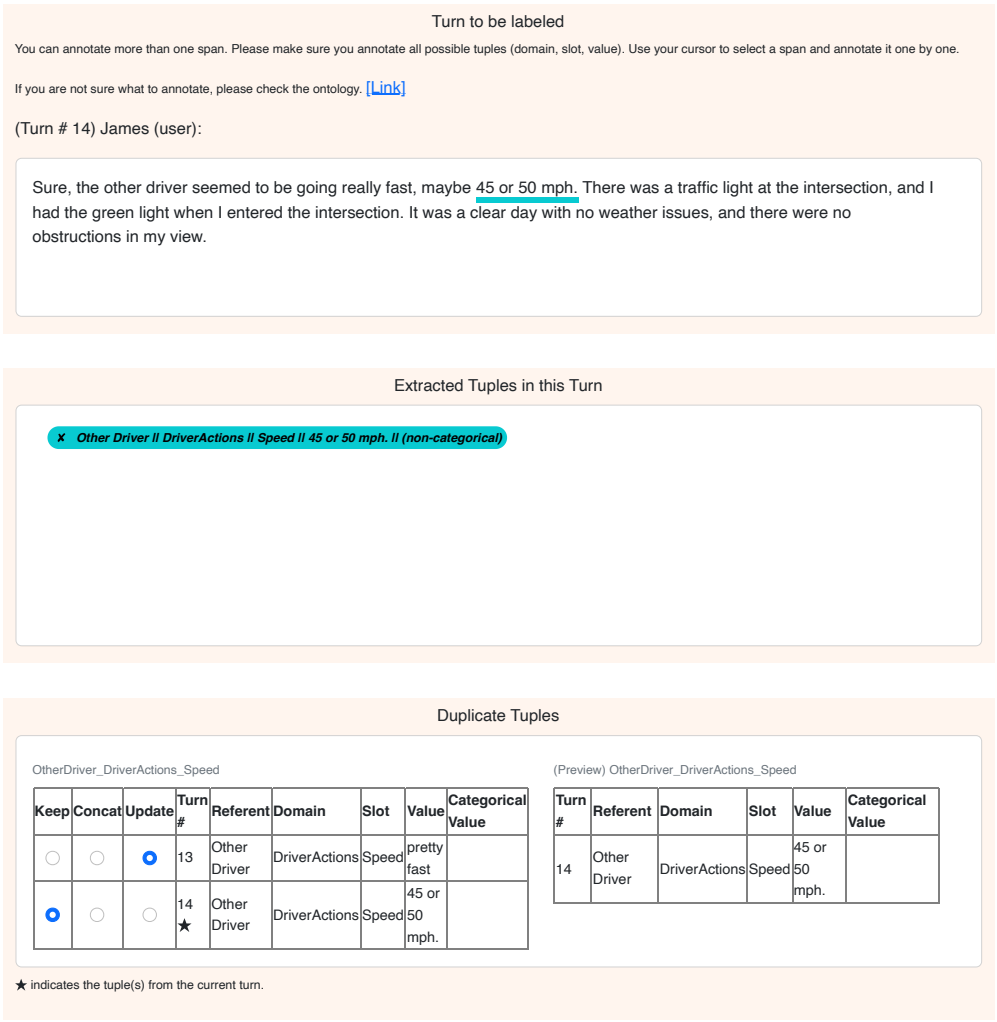


Figure 6: A human reviewer selects a span and label it. If there exists a duplicate label, they are prompted to resolve the conflict by selecting to update (as shown), concat, or keep multiple labels.

### E.5 DIALGEN-AIC dialogues

In Tables 11–13, we show sample dialogues from DIALGEN-AIC.

| | |
|---|---|
| Agent | Thank you for calling DialGen Insurance! This is Greta. How may I help you today? |
| User | Hello. This is Suzanne. I am calling for a car accident. |
| Agent | Alright, Suzanne. First, I need to verify some information. May I please have your full name and policy number? |
| User | Yes, my name is Suzanne Luna and my policy number is 1234567890. |
| Agent | Thank you, Suzanne. And can you please provide me with your phone number and address? |
| User | Yes, of course. My phone number is (123) 456-7890, and my address is 123 Main Street. |
| Agent | Thank you, Suzanne. Now, could you please describe the accident for me? |
| User | Certainly. I was driving alone on my way to the library when I approached an intersection with a yield sign. I slowed down and checked for oncoming traffic but didn't see any cars, so I began to turn left. Suddenly, I heard a loud crunching sound, and my car jerked to a stop. I got out of my car to assess the damage and saw that a black Nissan Maxima had hit the back of my car. |
| Agent | Thank you for letting me know, Suzanne. Do you know the name of the street you were on when the accident happened? |
| User | Sure, the accident happened at the intersection of Main Street and Broad Avenue. |
| Agent | And I believe you said you were driving alone, is that correct? |
| User | Yes, I was alone in my car. |
| Agent | Were there any witnesses at the scene who provided their contact information? |
| User | Yes, there were several witnesses at the scene who provided their contact information, but I didn't get their names. |
| Agent | Do you remember the date and time of the accident? |
| User | Yes, the accident occurred a few nights ago at around 11 pm. |
| Agent | Do you know the specific day? |
| User | Let's see, it would have been Wednesday–no, Thursday night. |
| Agent | Thank you, Suzanne. Can you tell me if your airbags deployed during the accident? |
| User | No, the airbags did not deploy. |
| Agent | Alright, and can you describe the traffic conditions at the time of the accident? Was there a traffic light at the intersection? |
| User | Yes, there was a traffic light and a yield sign. |
| Agent | I see. Did you notice anything else about the road or lanes that might have contributed to the accident? |
| User | No, nothing else that I recall. It all happened so fast. |
| Agent | Do you remember if you were going at a particular speed when the accident happened? |
| User | No, I was not speeding. I don't think the other driver was either. |
| Agent | Alright. Can you tell me about any damage your car sustained? |
| User | Yes, my car sustained significant damage on the left side. |
| Agent | I'm sorry to hear that. Do you know if a police report was filed? |
| User | Yes, the Houston Police Department took our statements, and the report number is 1234567890. |
| Agent | Do you know if either of you received a citation or ticket as a result of the accident? |
| User | Yes, the other driver, Homer Shepherd, received a citation. |
| Agent | Alright, thank you for letting me know. And can you provide me with the make and model of your car? |
| User | Yes, my car is a brown sedan, a Toyota Corolla. |
| Agent | Thank you, Suzanne. And can I ask if you sustained any injuries as a result of the accident? |
| User | Yes, unfortunately, I did sustain some injuries. I needed a CT scan at the hospital after the accident. |
| Agent | Can you tell me where you were injured, Suzanne? |
| User | My neck and back. |
| Agent | What was the diagnosis at the hospital? |
| User | The diagnosis was muscle strain and some minor bruising. |
| Agent | I'm sorry to hear that, Suzanne. How are you feeling now? |
| User | I'm doing a bit better, thank you for asking. |
| Agent | That's good to hear. Now, let me explain your coverage options to you. |
| User | Okay, thank you. |
| Agent | First, let me go over the details of your policy with you. You have liability insurance, which covers bodily injuries and property damage to others if you are at fault in an accident, although it looks like that's not applicable here. You also have collision coverage, which pays for damages to your vehicle in the event of an accident. |
| User | Okay, got it. |
| Agent | Additionally, you have personal injury protection, which will cover your medical expenses related to injuries sustained in the accident, regardless of who was at fault. However, please note that there may be limits on the amount of coverage provided by this policy. |
| User | Okay, thank you for explaining that to me. |
| Agent | Suzanne, can you tell me if your car was towed from the scene of the accident? |
| User | No, it wasn't. I was able to drive it home myself. |
| Agent | Okay, thank you for letting me know. I will just note that in the report. That should be everything for the preliminary details. Is there anything else you'd like to add, Suzanne? |
| User | No, that's all. Thank you for your help, Greta. |
| Agent | You're welcome. We will process your claim and be in touch with you soon about next steps. |
| User | Okay, thank you so much. |
| Agent | Have a good day, Suzanne. |
| User | You too, Greta. Bye. |

Table 11: Sample DIALGEN-AIC dialogue 1.

| | |
|---|---|
| Agent | Thank you for calling DialGen Insurance! This is Loan. How may I help you today? |
| User | Hello. This is Joshua. I am calling for a car accident. |
| Agent | I'm sorry to hear that you were in an accident. What happened? |
| User | It was last Monday morning in a tunnel. There were several cars involved, and my car was hit from the back by a yellow Dodge Ram. I was injured and had to be taken to the hospital by ambulance. |
| Agent | Oh my, I'm sorry to hear that. Did you have any passengers in your car? |
| User | No, I was the only one in the car. |
| Agent | Do you know how many cars were involved total? |
| User | I think there were about four cars involved. |
| Agent | Okay, thank you for that information, Joshua. Can you describe the damage to your car? |
| User | The back was heavily damaged, and my car is undrivable. |
| Agent | Was your car towed from the scene? |
| User | Yes, it was. Almost all of them had to be. |
| Agent | Did the police come to the scene of the accident? |
| User | Yes, they did. They took statements from witnesses, and they also created an accident report that documented all involved parties' details. |
| Agent | Great, do you happen to have the police report number and the name of the police department? |
| User | Yes, I have them right here. The police report number is 12345678, and it was the Philadelphia Police Department. |
| Agent | Thank you, Joshua. Was anyone cited or received a ticket at the scene? |
| User | No, the police report stated that no party was cited. |
| Agent | Okay, thank you for letting me know. Can you describe the traffic conditions at the time of the accident? |
| User | Traffic was flowing smoothly in the three-lane road. There was a car that stopped in the lane to my right, and the car behind them swerved into my lane. |
| Agent | Did you notice any traffic signals or signs that may have contributed to the accident? |
| User | No, there weren't any traffic signals or signs at all in the tunnel. |
| Agent | I see. Can you describe your car's make and model? What year was it made? And what color was it? |
| User | It's a white sedan, a 2018 Honda Accord. |
| Agent | Thank you for that information, Joshua. Were there any witnesses to the accident? |
| User | Yes, there were several people who saw the accident happen. Some good Samaritans helped me after the accident and called 911. |
| Agent | That's good to hear. Now, can you tell me about your injuries? What kind of medical treatment did you receive? |
| User | I dislocated my shoulder. They performed a CT scan at the hospital to ensure that there were no internal injuries. |
| Agent | One more thing, Joshua. Can you remind me of the exact date and time of the accident? |
| User | It was on Monday morning, around 8:30 am. |
| Agent | Okay, just to confirm, that would be the 22nd, correct? |
| User | Oh, wait. I think I may have remembered it wrong. It was actually last Tuesday. |
| Agent | Thank you for clarifying the date, Joshua. Can you also tell me how fast were you driving when the accident occurred? |
| User | I was driving around 35 miles per hour. |
| Agent | Thank you for that information, Joshua. Do you have the contact information for any of the other drivers? |
| User | Yes, I got Steve Woods' phone number. She was driving the yellow Dodge Ram that hit my car. |
| Agent | Joshua, can you confirm how the accident occurred from your perspective? I'm a little unclear on some of the details. |
| User | Sure, I was driving in the middle lane and noticed a stopped car in the other lane. The yellow Dodge Ram swerved into my lane to avoid it. I couldn't change lanes because the other lane was bumper to bumper. The Dodge Ram hit my back bumper, bounced off, and hit the car on the right side of me. After that, several cars collided with each other. My car spun around and hit the tunnel wall, damaging the back. |
| Agent | Thank you for explaining that, Joshua. Just to confirm, were there any traffic cameras at the scene that may have captured the accident? |
| User | No, I don't believe there were any traffic cameras. |
| Agent | Thank you for providing all of that information, Joshua. Can I ask you to confirm some information about yourself? |
| User | Sure, go ahead. |
| Agent | Can I have your full name? |
| User | Joshua Walters. That's spelled W-A-L-T-E-R-S. |
| Agent | And can I have you provide your phone number and email address please? |
| User | Yes, my phone number is 555-1234, and my address is 123 Main Street. |
| Agent | Great, thank you for providing that information. Now it looks like your insurance coverage includes liability insurance, collision coverage, rental car coverage, and gap insurance. |
| User | Okay. |
| Agent | Do you understand how each of those coverages works? |
| User | Not really, can you explain it to me? |
| Agent | Sure, liability insurance covers any damages or injuries that you may cause to another person or their property. Collision coverage pays for damages to your own vehicle if you're in an accident. Rental car coverage pays for a rental car while your vehicle is being repaired. And gap insurance covers the difference between the actual cash value of your car and the balance remaining on your car loan if your car is totaled. |
| User | Okay, I think I understand now. Thank you for explaining that to me. |
| Agent | Do you have any other questions for me, Joshua? |
| User | No, I don't. Thank you for your help. |
| Agent | Before we end this call, do you happen to have any other details about the accident that you would like to add? |
| User | No, I think that covers everything. |
| Agent | Okay, thank you for providing all the information, Joshua. We will take care of everything from here. |
| User | Thank you so much for your help. |
| Agent | Of course! And please don't hesitate to reach out if you have any questions or concerns. |
| User | I won't. Thanks again. |

Table 12: Sample DIALGEN-AIC dialogue 2.

| | |
|---|---|
| Agent | Thank you for calling DialGen Insurance! This is William. How may I help you today? |
| User | Hello. This is Roberta. I am calling for a car accident. |
| Agent | I am sorry to hear that, Roberta. Before we get started, can I get your permission to record this call for quality and training purposes? |
| User | Yes, you can record the call. |
| Agent | Thank you. May I have your full name and policy number, please? |
| User | My name is Roberta Oliver, and my policy number is XXXXXXXX. |
| Agent | Got it. Can you please describe the accident to me? |
| User | Sure, I was driving my gold rideshare car on the highway after having dinner at a restaurant. I was changing lanes when Shane Scott's gray Chevrolet Camaro cut me off, and I had to slam on my brakes to avoid collision. Unfortunately, it was too late, and Shane's car crashed into mine. |
| Agent | Okay, thank you for the detailed account. Were there any witnesses who saw this happen? Any traffic controls? |
| User | No, I didn't see any traffic controls around. I'm not sure about any witnesses. Oh, I guess there were the passengers in Shane's car, but they were too shaken up to give their statements to the police. |
| Agent | Alright. How many passengers were in each car? |
| User | Shane had three passengers in her car. I was alone in mine. |
| Agent | Thank you for that information, Roberta. Can you provide me with the location details of the accident as well as the date and time it occurred? |
| User | It was May 15th at around 4 in the afternoon. The accident happened on the highway near exit 45B. |
| Agent | Thank you for sharing that information, Roberta. I forgot to ask earlier, what year is your car? |
| User | My car is a 2012 model. |
| Agent | Great, thanks for letting me know. Can you describe the traffic conditions at the time of the accident? |
| User | It was a beautiful day, and the traffic on the highway was moving at a steady pace. There were four lanes, and we were both in the second lane from the left. |
| Agent | Alright, I see. Before we proceed further, I want to let you know that I understand how stressful this situation can be. I want you to know that I am here to guide you through the process and make everything as clear and easy as possible. How are you feeling? |
| User | Honestly, I'm feeling pretty overwhelmed right now. My head has been hurting since the accident, and I'm worried about how much this is all going to cost. |
| Agent | That's perfectly understandable, Roberta. Just take a deep breath and try to relax. It's good that you're taking steps towards resolving this by calling us today. Let's move forward together, okay? |
| User | Okay, thank you. |
| Agent | Now you mentioned your head has been hurting since the accident. Did you injure your head during the crash? |
| User | Yeah, I hit my head on the steering wheel. Since then, I've been having constant headaches. It's been really difficult to focus on everyday tasks. |
| Agent | I'm sorry to hear that. Have you seen a doctor yet? |
| User | Yes, I went to the hospital after the accident. They gave me a CT scan which revealed that I had a minor concussion. |
| Agent | I'm sorry to hear that. Did they prescribe any treatment or medication? |
| User | Not really, other than rest and avoiding physical activities. They okayed me to go back home immediately, but I needed to have my husband check on me every few hours to make sure everything was fine that first night. |
| Agent | Have you been back to the hospital since to follow up on the headaches? |
| User | No, but I did call my doctor to ask her about it. She said that headaches are normal for the first couple of months after a concussion, but to go back if they get worse. |
| Agent | I see. Thank you for telling me that, Roberta, and I hope the headaches get better soon. Just a few more questions if you'll bear with me. Can you tell me which part of your car was damaged in the accident? |
| User | The front left side of my car was damaged. The back right side of Shane's car as well. |
| Agent | Thank you for that information. Now I understand that it can be frustrating when there are no witnesses to corroborate your story. However, do you have any evidence of the accident? Perhaps photos of the damage or the police report? |
| User | Yes, the police came to file a report. I have a copy of it at home. I also took some photos of the damage to my car and Shane's car. |
| Agent | Great, that will certainly help. Can you please send those photos over to our team? I can provide you with an email address where you can send them. |
| User | Sure, that would be helpful. What's the email address? |
| Agent | The email is claims@DialGen Insurance.com. Please put your full name and policy number in the subject line and attach the photos in the email body. |
| User | Okay, thanks. I will send them over as soon as possible. |
| Agent | Perfect. Is there anything else I can assist you with today, Roberta? |
| User | Yes, I was wondering about the insurance claim process. How long does it usually take to get a resolution? |
| Agent | It depends on a few factors, such as the complexity of the case and how much evidence we have. Our team will carefully review your claim and reach out to you within a few business days with a resolution. |
| User | Okay, that's good to know. And what about rental cars or any other expenses related to the accident? |
| Agent | We can certainly help you out with that if you need it. Our team can set up rental cars if necessary, and we will do everything we can to make sure you're not paying out of pocket for any expenses related to the accident. Will you be needing a rental car? |
| User | No, I don't think so. |
| Agent | Alright, no problem. If you do end up needing a rental car, feel free to let us know. We're here to help in any way we can. |
| User | Thanks, I appreciate it. |
| Agent | Of course, Roberta. Is there anything else I can assist you with today? |
| User | No, that's all for now. Thanks for your help, William. |
| Agent | It was my pleasure, Roberta. Take care and have a great day! |
| User | You too. |

Table 13: Sample DIALGEN-AIC dialogue 3.

## F Additional analysis

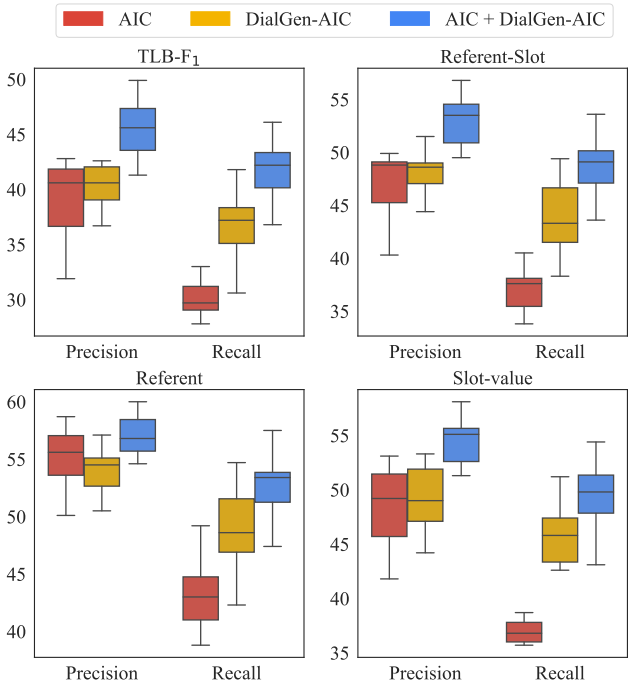


Figure 7: TLB and three diagnostic scores for precision and recall ($m_{\text{R}}$, $m_{\text{RS}}$, and $m_{\text{SV}}$) for the T5-SC model on AIC test set.

Figure 7 provides the TLB precision and recall results for the full state updates and different diagnostic scores (referent only, referent-slot, and slot-value). Consistent with the CB results, the biggest benefit of incorporating DIALGEN-AIC is improved recall. While referent, slot, and value all improve, the greatest improvement is in slot values.

## G License of artifacts

The license of code for (Wolf et al., 2020) is Apache license version 2.0. The license of code for Faker and Gender-guesser is MIT and GPLv3 License, respectively. The terms for use of our artifacts will be included in our released package.