S3-DST: Structured Open-Domain Dialogue Segmentation and State Tracking in the Era of LLMs

Sarkar Snigdha Sarathi Das[1,†,‡], **Chirag Shah**[2,‡], **Mengting Wan**[3], **Jennifer Neville**[3], **Longqi Yang**[3], **Reid Andersen**[3], **Georg Buscher**[3], **Tara Safavi**[3,†]
[1]Pennsylvania State University, [2]University of Washington, [3]Microsoft
[†]Corresponding authors: sfd5525@psu.edu, tarasafavi@microsoft.com
[‡]Work done at Microsoft, USA

Abstract

Traditional Dialogue State Tracking (DST) has focused on tracking preferences and intents in conversations centered around *specific* tasks (e.g. booking services). These conventional systems assume a relatively restricted conversation flow in which each turn gradually offers new information. However, advancements in Large Language Models (LLMs) have ushered in more versatile *open-domain* chat systems in which extended dialogue sessions encompassing numerous tasks and topics are common—in turn requiring new conversational tracking tools in order to successfully orchestrate such systems. Addressing these challenges, we introduce a novel approach combining dialogue segmentation and state tracking within open-domain dialogues, tailored for zero-shot applications appropriate to a true open-domain dialogue system. Our proposed method S3-DST employs a unique structured prompting technique called *Pre-Analytical Recollection*, a novel grounding mechanism we designed for improving long context tracking. Tested on proprietary anonymized open-domain dialogue datasets as well as publicly available DST and segmentation datasets, S3-DST consistently outperforms the state-of-the-art, showcasing its effectiveness and adaptability state tracking in the next wave of LLM-based chat systems. We also release S3-DST annotations with GPT-4 on a curated subset of LMSYS-Chat-1M (Zheng et al., 2023) to be used as a testbed to fuel research in this direction.

1 Introduction

The advent of open-domain Large Language Model (LLM)-based chat systems like ChatGPT and Bing Chat has ushered in a new age of dialogue systems. Previously, dialogue systems were relatively constrained in their scope and abilities, typically confined to either narrow task-oriented conversations or social chitchat (Gao et al., 2018). By contrast, LLM-based chat systems are remarkable because they can converse fluidly with users over a seemingly infinite range of topics, and can accomplish many user tasks out-of-the-box that previously required specialized systems, like code generation, question answering, and more.

```
User: Please find an annotated Bibliography for "The Road"
AI: Here is an example annotated bibliography for …
[Call Search API]
…
User: Thank you! That helps a lot.
AI: You are very welcome. Anything else I can help you with?
…
User: Cover photo of "The Road"
AI: Here is the cover photo of "The Road" I've found:
[Call Image Search API]

User: That's not what I am looking for.
AI: Oh, I am sorry. Maybe you can tell me more about it?

User: Umm, I want a cover photo in the style of Van Gogh.
AI: I see. I'll try to create that!
[Call Generation API]
```

Figure 1: A single intent may span several turns in open-domain conversation, and a single conversation may contain multiple intents: A synthetic dialogue inspired by anonymized Bing Chat logs. Different user intents (searching for an annotated bibliography, social chitchat, generation of a cover photo in target style) are highlighted by different colors.

While LLM-based chat systems have significantly changed the landscape of human-AI dialogue, the enhanced fluency and unrestricted nature also significantly increase the challenges associated with the analysis and tagging of dialogues, due to their closer resemblance to fluid real-world conversation. Conventional chat agents designed for specific tasks interact with users in a predictable sequence, extracting required necessary information (state values) turn-by-turn to properly serve user requests (Williams et al., 2016; Budzianowski et al., 2018). However, the traditional method of turn-by-turn dialogue state tracking (DST), although effective for previous domain-specific tra-

Findings of the Association for Computational Linguistics: ACL 2024, pages 14996–15014
August 11-16, 2024 ©2024 Association for Computational Linguistics

ditional chatbots, does not adequately address the complexities of real-world, open-domain dialogues. These conversations often involve a considerable amount of back-and-forth interaction between the user and agent, and the context may shift several times within a single dialogue, moving between unrelated topics and/or intents. This dynamic is illustrated in Figure 1, where open-domain dialogues facilitated by LLM-based chat systems enable users to pursue various intents within a single conversation, thereby creating distinct topical and/or goal-oriented segments. As depicted, certain segments necessitate multiple exchanges between the user and the agent to accurately determine the user's intent, which then informs the orchestration layer and the selection of relevant APIs (e.g. web search, image or code generation, etc.) to satisfy the user's request. A turn-by-turn tracking scheme is thus insufficient to accurately capture dialogue state changes across multi-turn segments.

In order to tackle this, we propose to track both **segments** and **states** in open-domain dialogue: *Segmentation* helps us identify boundaries that mark the start and end of contextually cohesive conversation "units," whereas *states* are the intent variables of interest we wish to track, applied *per segment*. Beyond bringing DST into the era of open-domain conversation and LLMs by joint tracking of segments and states, we introduce LLM-based *solutions* for open-domain DST. Assuming a zero-shot setting for dialogue tagging, which is realistic due to the cost of labeling, we introduce **S3-DST**, a structured prompting approach for open-domain DST. Within S3-DST we propose a novel Pre-Analytical Recollection (PAR) prompting strategy that grounds each output state prediction on the content of the corresponding dialogue turn, thereby helping the LLM track long dialogue context without forgetting or hallucination.

We evaluate S3-DST on a fully anonymized open-domain dialogue dataset collected from Microsoft's Bing Chat system, alongside public DST and segmentation benchmarks.[1] S3-DST outperforms comparable baselines significantly across all benchmarks, indicating its potential as an initial framework for advancing research in open-domain dialogue modeling. Moreover, we release S3-DST annotations with GPT-4 on a curated subset of LMSYS-Chat-1M (Zheng et al., 2023), which we believe will be an excellent starting point for understanding and benchmarking open-domain DST.

[1]The use of Bing Chat logs is in compliance with the terms of use of Bing Chat.

2 Problem Definition

Informally, the goal of traditional DST is to predict the dialogue state y_t given a sequence of user and agent utterance turns $C_t = [U_1, A_1, \dots, U_t, A_t]$.[2] The state y_t consists of a set of slot-value pairs, where slots correspond to intent attributes in a particular application domain (e.g., "restaurant-name", "hotel-address") and values correspond to predefined categorical options or unconstrained text (Budzianowski et al., 2018).

However, as we have previously discussed, a single open-domain conversation will often consist of multiple potentially unrelated intents across a variety of topics. Indeed, according to a preliminary analysis on 10K anonymized Bing Chat conversations, we estimate that over 50% of conversations display multiple user intents and over 90% of conversations contain discussion of multiple topics. Therefore, we propose to merge dialogue segmentation, which aims to find contextually cohesive "units" of dialogue within a larger conversation, with dialogue state tracking. In particular, we perform state tracking at the *segment* level, where the goal is to label each segment with the slots and values of interest, such that multiple segments within a conversation may have diverging or conflicting state values, reflecting the true variety of open-domain chat.

In the rest of this section, we define segmentation and state, and finally formalize the joint task.

2.1 Segment

Following previous work in dialogue topic segmentation (Xing and Carenini, 2021; Xia et al., 2022; Gao et al., 2023), we define **dialogue segments** as contiguous subsequences of C_t in which all user and agent utterances are topically related. Formally, let $B_t = [b_1, \dots, b_{t-1}]$ indicate the boundary indices between adjacent user-agent utterance pairs in C_t. The output of segmentation is a set of boundary indices $B_k \subseteq B_t$, where k represents the number of boundaries determined by the segmentation algorithm and the span $[U_m, A_m, \dots U_n, A_n]$ represents the contiguous segment between boundaries b_m and b_n, where $m \in [1, t-1]$ and $n \in [m, t-1]$.

[2]Note that in current LLM-based chat systems, users may issue multiple utterances before a single agent response is issued. In these (infrequent) cases, we group all user utterances prior to the agent response into a single utterance.

2.2 Segment state

Typically, dialogue state tracking methods extract new elements of state at each turn (Hu et al., 2022). However, this is because DST evaluation benchmarks make the relatively narrow assumption that users provide new and relevant elements of intent at each turn, and that intents build upon or complement each other but do not fundamentally change or conflict throughout the conversation. As we have previously discussed, open-domain dialogue exhibits far more varied characteristics, and multi-intent and/or multi-domain conversations are relatively common.

We therefore propose to extract state at the segment rather than turn level. We define the segment-level state as $\{S_{m:n} = (s^{(i)}_{m:n}, v^{(i)}_{m:n}), i = 1 \dots N_{m:n}\}$, where $s^{(i)}_{m:n}$ refers to the i-th slot applied to the segment from boundaries b_m to b_n, $v^{(i)}_{m:n}$ refers to the slot's corresponding value, and $N_{m:n}$ refers to the total number of slots to applied to this segment.

Any schema of slot-value pairs is valid under this setting; in our particular Bing Chat state schema we are interested in tracking *Segment Intent* and *Segment Domain*, whereas we adhere to the provided state schemas in public datasets. We describe our detailed schema for Bing Chat in § 4.1 and Appendix B.

2.3 Problem statement

Having defined segments and per-segment state, we are equipped to state our full definition of open-domain DST. Given a sequence of user-agent utterance pairs $C_t = [U_1, A_1, \dots, U_t, A_t]$, we define the goal of open-domain dialogue state tracking as jointly predicting

$$y_t = B_k \cup \{S_{m:n}\,;\, \forall (b_m, b_n) \in B_k\}, \qquad (1)$$

where $B_k \subseteq B_t$ refers to the segment boundary indices described earlier and $S_{m:n}$ refers to the segment state between boundaries b_m and b_n, consisting of N arbitrary slot-value pairs:

$$S_{m:n} = \{(s^{(i)}_{m:n}, v^{(i)}_{m:n}), i = 1 \dots N_{m:n}\}. \qquad (2)$$

3 Prompting Strategies

As discussed previously, real-world dialogues often span multiple topics and intents. Previous studies (Hu et al., 2022) aimed at disassociating individual dialogue turns and processing them one by one for tracking dialogue state changes, which worked reasonably well in task-oriented dialogues within predefined narrow domains. However, real-world dialogues commonly require multiple turns to adequately comprehend contextual nuances, which is a challenge because Transformers still struggle when processing lengthy input contexts, particularly in the middle (Liu et al., 2023). To address these difficulties, we propose a novel turn-by-turn prompting technique that gives structure to inputs and outputs while accurately preserving the context in the process. We discuss our approach below and summarize it in Figure 2:

3.1 Structured Outputs and Inputs

Structured Output Our goal is a set of labels per dialogue turn representing the segment boundaries (binary labels) and state values (categorical labels or open text). To provide a flexible yet structured format to the LLM's output, we propose to instruct it to generate outputs in a hierarchical XML format. We see XML as advantageous because it provides code-like structure to the DST task, which has been shown to greatly improve performance compared to plain-text outputs, while still being extensible and flexible compared to more rigid output formats like SQL (Hu et al., 2022).

Our approach uses an XML format in which each turn from 1 to t comprises an XML tree <T{id}>...</T{id}> and several nested XML tags within it. The labels of these nested tags (e.g. <segment_start>...</segment_-start>, <intent>...</intent>, and <domain>...</domain> in Figure 2(ii)) represent the segment boundaries and slots of interest, and each value between opening and closing tags represent the model's inferred value.

This strategy is beneficial for two reasons: (i) Due to a bounded and well-defined structured formatting, generated outputs are more likely to be aligned with labeling instructions than free-form texts; and (ii) well-formed structured output formats are easier to parse, thus reducing postprocessing requirements.

Structured Input In terms of prompting an LLM to analyze a dialogue history, we observe that a plain-text format makes it difficult for the LLM to refer back to, and reason about, multiple conversational turns. To handle this challenge, consistent with the output format, we propose a structured conversational input format, where each di-

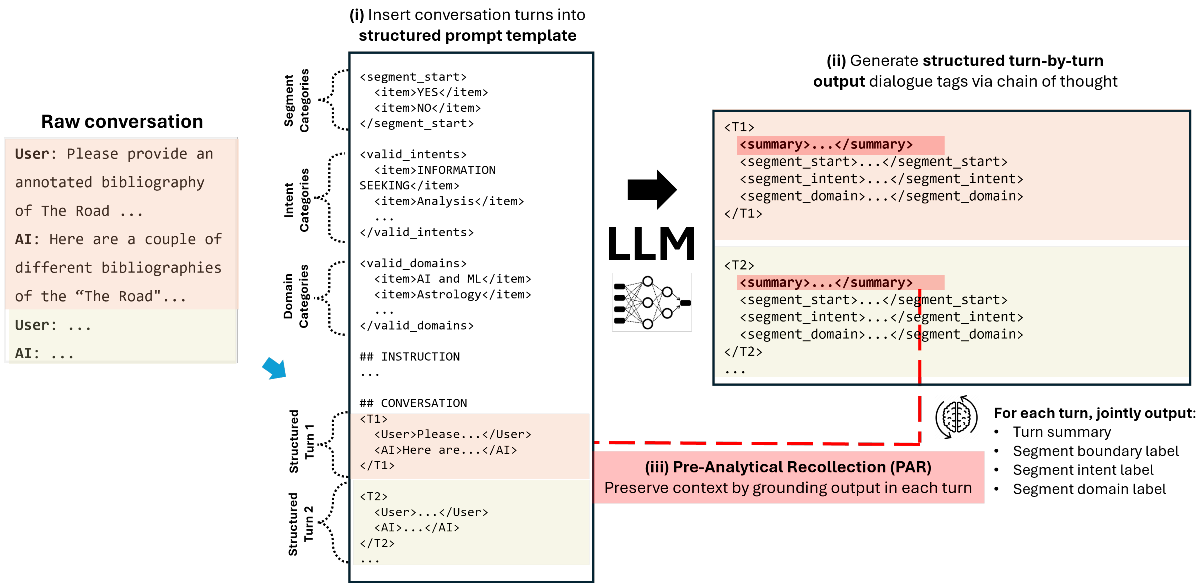


Figure 2: Prompt flow of S3-DST. Given a **raw conversation**, (i) we convert it into a hierarchical XML-structured representation (**Structured Turn**) and insert it into a similarly **structured prompt template**. The target valid **Segment/Intent/Domain Categories** are specified at the beginning of the prompt. We pass the prompt through the LLM and (ii) obtain a hierarchical XML-structured **turn-by-turn output**, where each turn contains (iii) a **Pre-Analytical Recollection (PAR)** grounding reference to the conversation alongside the desired segmentation and state label predictions. For the actual wording of the prompt, please refer to Appendix A.1

alogue history is formed into a hierarchical XML format in which conversational turns are marked with a turn ID number <T{id}>...</T{id}> numbered from 1 to t and each conversational turn consists of nested user and agent turns marked with appropriate XML tags (<user>...</user> and <agent>...</agent>). This input scheme helps the LLM accurately refer back to the input turn and maintain coherence for long dialogue contexts.

Consistent with this XML-tagged input format, we also format all the valid segment and state category labels into an XML-formatted list using the following structure: <valid_category_name> <item>{label name}</item> <description>{description of label, if available}</description> </valid_category_name>. Empirically, this structured input and prompt formatting help constrain the LLM generation to follow the labeling instructions. Figure 2(i) shows this format where each valid segment boundary and state category are first staged in an XML-formatted list and subsequently input dialogue is staged in a hierarchical configuration.

3.2 Pre-Analytical Recollection (PAR)

As previously discussed, open-domain dialogues may be long and highly variable in conversation flow. Therefore, it is crucial to ensure that the LLM can accurately monitor the evolving dialogue context without forgetting or hallucination. To this end, we propose Pre-Analytical Recollection (PAR), a grounding strategy for turn-by-turn prompting that instructs the LLM to first summarize the turn using <summary>...</summary> tags in 3 sentences or fewer before providing the segment and state values. PAR is inspired by chain-of-thought prompting (Wei et al., 2022), as it is a technique for generating relevant intermediary outputs in order to improve reasoning accuracy. However, unlike chain-of-thought, PAR is also a grounding technique that provides references from the model's output directly to the conversation. Figure 2(ii) demonstrates how PAR prompts the LLM to refer back to the content of each conversational turn before analyzing it to infer the conversational states. We find that this strategy greatly improves the consistency and accuracy of state tracking, as we will demonstrate in the experiments.

3.3 Final Prompt Configuration

The final prompt flow of S3-DST is provided in Figure 2. Given a raw conversation and a predefined set of segment and state labels, we insert the labels into a structured prompt template and format the conversation in a hierarchical XML-structured representation. We pass the prompt through the

Table 1: Evaluation test set statistics.

	# Convs	# Turns	# segments/conv (avg.)
Bing Chat	334	2308	1.51
MWOZ 2.1	1,000	7368	-
MWOZ 2.4	1,000	7368	-
DialSeg711	711	19350	3.87
LMSYS-Chat-Split	5,100	24046	1.81

LLM, instructing it to follow PAR before jointly generating the hierarchical turn-by-turn segmentation and state labels applied per segment. The full text of our prompt is provided in Appendix A.1.

4 Experiments

We conduct comprehensive evaluations across multiple datasets. We primarily evaluate our approach on fully anonymized Bing Chat logs annotated by domain experts. Additionally, we evaluate S3-DST on the standard task-oriented DST and segmentation tasks using public benchmark datasets MultiWOZ (Budzianowski et al., 2018) and DialSeg711 (Xu et al., 2021) respectively.

Finally, we also construct a curated set of conversations from the recently open-sourced LMSYS-Chat-1M (Zheng et al., 2023) dataset, which is a similar open-domain human-LLM dialogue dataset. We apply S3-DST with GPT-4 and open-source the results, which we believe can be highly useful as a starting point for open-domain dialogue analysis and benchmarking. We discuss the LMSYS-Chat-Split dataset in more detail in Appendix D. A detailed description of the other datasets is provided below, alongside the dataset statistics in Table 1.

4.1 Internal Human-LLM Dialogue Dataset

In order to evaluate the efficacy of our approach on real-world open-domain human-LLM conversations, we collected anonymized chat log data from Microsoft's Bing Chat system, an LLM chat interface backed by the Bing search engine. As mentioned previously, a large-scale preliminary analysis of conversations logged by the system show that over 50% of conversations display multiple intent and over 90% display multiple domains; while we are limited in terms of sharing dataset specifics due to privacy constraints, we share distributional plots and representative conversations from the open-domain LMSYS-Chat-1M dataset in Appendix D, which we find are qualitatively similar to those observed in our internal dataset.

Benchmark construction We sample 484 English conversations conducted on Bing Chat between April 5, 2023 to April 30, 2023 via two approaches: (i) Random and (ii) "Long" conversations of 5 or more turns only. We balance these two approaches 50/50. Since we operate under a zero-shot assumption, we do not need any training data. Therefore, we hold out 150 conversations for development and the remaining 334 for testing.

Annotation To obtain ground-truth labels for evaluation, we gathered human annotations for segment and state. We recruited three in-house annotators not affiliated with this paper. For each turn, we instructed annotators to provide binary **IsSegmentBoundary** labels, categorical **SegmentIntent** labels, and categorical **SegmentDomain** labels. We instructed annotators to mark a segment boundary when no topical relation between a turn and its preceding context could be identified. For intent and domain, we used taxonomies developed in-house for the Bing Chat system consisting of 4 intents (Information Seeking, Analysis, Creation, and Open-Ended Discovery) and 49 domains (see Appendix B.1 for the full list).[3] Appendix B provides further details on the annotation scheme and setup. To ensure inter-annotator agreement before labeling the full dataset, we first gathered annotations on a set of 10 randomly selected conversations (68 turns total) and computed Fleiss' kappa (Fleiss, 1971) per label type. We observed a Fleiss kappa of $\kappa = 0.83$ for **IsSegmentBoundary**, $\kappa = 0.74$ for **SegmentIntent**, and $\kappa = 0.88$ for **SegmentDomain**, all of which are considered high agreement on the Fleiss kappa scale.

4.2 Public Benchmarks

We are not aware of any existing public dialogue benchmarks reflective of the broadly open-domain Bing Chat data. Therefore, we resort to separate DST and segmentation evaluations on public benchmarks using three datasets.

MultiWOZ The MultiWOZ (MWOZ) multi-domain dialogue dataset (Budzianowski et al., 2018) is currently the most common DST benchmark. MWOZ is a task-oriented dataset consisting

[3] Note that in our case, intents are agnostic of domain, as we find that these four intents are universally applicable across all domains in Bing Chat. However, in DST, intents are typically domain-specific; our approach can handle this easily by simply prefixing each intent with its relevant domain, as has been done in previous work.

of 1K test dialogues. We use two updated versions of the original: MWOZ 2.1 (Eric et al., 2019) and 2.4 (Ye et al., 2021). The latter is considered the "cleanest" version of MWOZ, while the former has been used more frequently in the literature.

DialSeg711 The DialSeg711 benchmark was introduced by (Xu et al., 2021) and has been used frequently in recent dialogue segmentation research. It is an English dataset in which 711 multi-segment dialogues are constructed by joining dialogues from existing task-oriented dialogue corpora.

4.3 Baselines

As baselines we consider zero-shot LLM prompts only, for a fair comparison to S3-DST. We discuss the baselines and their considerations below for different datasets. All original prompts are provided in Appendix A. We set a maximum of 1500 output tokens per LLM call with a temperature of zero.

Bing Chat In this dataset, we consider **IC-DST** as our primary baseline, which is a zero-shot version of the prompting strategy introduced by (Hu et al., 2022), heavily adapted for open-domain dialogue setting to jointly track segment and dialogue states. The **TBT-DST** baseline is a version of S3-DST that does not include segmentation instructions and obtains intent and domain labels on a turn-by-turn basis using our S3-DST prompt configuration. Moreover, to analyze the importance of two key aspects of our prompt, PAR and XML-structured formatting, we also consider two ablations of S3-DST: **No PAR** refers to a S3-DST prompt without the PAR instructions, and **Unstructured input** refers to a S3-DST prompt that formats all instructions and dialogue using plain text rather than XML. We use GPT4 as the backbone LLM for all prompts in our main experiments.

MWOZ For MWOZ task-oriented dialogue state tracking dataset, we compare against **IC-DST using Codex-175B** as reported by Hu et al. (2022). We also reevaluate zero-shot **IC-DST with GPT-4** to account for the backbone model improvement in baseline performance. However, we also experimented with smaller open-source language models (**Llama2-70B-chat** (Touvron et al., 2023)) to assess their potential for zero-shot DST. We explain our findings in detail in Appendix C. Moreover, we measure the performance of D3ST (Zhao et al., 2022) where a T5-XXL(11B) model is finetuned on SGD dataset and then transferred to MultiWoz

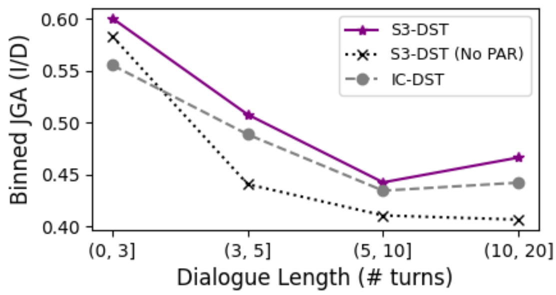


Figure 3: S3-DST outperforms baselines for dialogues of all lengths by emphasizing context tracking. We bin Bing Chat dialogues by length and plot JGA per bin. The large performance degradation of both baselines as the dialogue length increases confirms the importance of our PAR grounding strategy.

2.4 dataset. Also, we compare against the zero-shot **ChatGPT** performance on MWOZ 2.1 as reported by (Heck et al., 2023). Finally, for per-domain JGA performance comparison, we compare against IC-DST (Hu et al., 2022) and dialogue template modeling **(DM)** (Campagna et al., 2020) with TRADE (Wu et al., 2019) and SUMBT (Lee et al., 2019).

DialSeg711 We consider the unsupervised **TextTiling** (Hearst, 1997), **CSM** (Xing and Carenini, 2021), and **DialStart** (Gao et al., 2023) methods. We reprint all numbers from (Gao et al., 2023). Finally, we use our **IC-DST** baseline prompted to elicit segmentation labels in the same SQL output format as the original IC-DST (Hu et al., 2022).

4.4 Metrics

For state tracking, we consider **Joint Goal Accuracy (JGA)**, which measures the proportion of turns for which all state values are correctly inferred. For Bing Chat, we report JGA with just intent and domain (I/D) as these are the true state values of interest, as well as JGA with segment, intent, and domain accuracy (S/I/D) for completeness. We also report segmentation, intent, and domain accuracy separately on Bing Chat to provide a sense of the current capabilities and limitations of LLMs on open-domain conversational data. For segmentation, we consider P_K and **WindowDiff** (Pevzner and Hearst, 2002), which are both error metrics (i.e., lower is better) that quantify the difference between predicted and ground-truth segment boundaries using an adjustable sliding window.

Table 2: S3-DST achieves state-of-the-art performance on state tracking over our internal Bing Chat benchmark. All prompts are run with GPT4.

	Individual accuracy			JGA	
	Segment	Intent	Domain	I/D	S/I/D
TBT-DST	-	0.6707	0.6221	0.4169	-
IC-DST	0.8567	0.7123	0.6049	0.4610	0.4387
S3-DST (No PAR)	0.8859	0.7173	0.6251	0.4377	0.4078
S3-DST (Unstructured input)	0.8810	0.7163	0.6307	0.4640	0.4331
S3-DST	**0.8992**	**0.7366**	**0.6429**	**0.4752**	**0.4504**

Table 3: S3-DST achieves state-of-the-art JGA compared to zero-shot LLM baselines on the public dialogue state tracking benchmarks MWoZ 2.1 + 2.4.

	JGA	
	MWOZ 2.1	MWOZ 2.4
IC-DST (Codex)	0.3534	0.3530
IC-DST (GPT4)	0.4045	0.4625
D3ST (T5-XXL)	-	0.2890
ChatGPT	0.3150	-
S3-DST	**0.4513**	**0.5327**

Table 4: Zero-shot per-domain comparison (JGA) on MWOZ 2.1.

	Per-domain JGA				
	attr.	hotel	rest.	taxi	train
IC-DST (Codex)	0.5997	0.4669	0.5728	0.7135	0.4937
IC-DST (GPT4)	**0.7177**	0.4872	0.6526	0.7781	0.5710
ChatGPT	0.5270	0.4200	0.5580	0.7090	0.6080
TRADE-DM	0.3490	0.2830	0.3590	0.6500	0.3740
SUMBT-DM	0.5280	0.3630	0.4530	0.6260	0.4670
S3-DST	0.6781	**0.5215**	**0.6713**	**0.8258**	**0.7027**

4.5 Results

Bing Chat As shown in Table 2, our S3-DST prompt achieves the highest performance across intent, domain, and JGA across turns. We make the following observations: First, TBT-DST, which does not explicitly perform segmentation, is by far our weakest baseline. We find that this is because without instructing the LLM to use the same intent and domain within a segment, the LLM tends to overindex on the content of the turn without considering the fuller preceding context. This leads to conflicting intent and domain labels between turns within a coherent single-topic dialogue.

Second, our adapted version of IC-DST is a very strong baseline. However, while IC-DST makes use of structured outputs, it does not have a corresponding structured input representation. We find that this hurts its performance in some cases, as hallucination of nonexistent turns is relatively more common compared to S3-DST.

Finally, the two ablations of S3-DST both underperform compared to S3-DST, confirming the importance of PAR and structured inputs that the LLM can refer back to during generation. Indeed, Figure 3, which plots the relationship between dialogue length and performance, shows that S3-DST avoids the steep degradation in performance of the no-PAR ablation as the dialogues get longer. For example, the no-PAR ablation performs comparably to S3-DST on conversations of 3 turns or fewer, but drops over 10 points JGA for conversations of 4 turns or more. These results in particular highlight the necessity of PAR for long dialogues.

Table 5: S3-DST achieves state-of-the-art performance on the public segmentation benchmark DialSeg711.

	$P_k(\downarrow)$	WindowDiff $(\downarrow)$
TextTiling	0.4044	0.4463
CSM	0.2430	0.2635
DialSTART	0.1786	0.1980
IC-DST	0.2889	0.2419
S3-DST	**0.0091**	**0.0081**

MWOZ Tables 3 and 4 provide MWOZ numbers in total and per-domain. S3-DST achieves state-of-the-art zero-shot JGA compared to strong LLMs by a large margin. Even our strongest zero-shot baseline, IC-DST (GPT4), has an absolute performance gap of nearly 5 points JGA on MWOZ 2.1 and 7 points on MWOZ 2.4. In nearly all individual domains, S3-DST outperforms IC-DST (GPT4), and some by a large margin, for example over 13 points JGA improvement on the train domain.

DialSeg711 Finally, Table 5 shows performance on DialSeg711. S3-DST achieves nearly zero error on this dataset, which we find unsurprising given that the dataset's construction. Specifically, DialSeg711 is constructed by joining dialogues about very different topics, which leads to very artificial and abrupt context shifts between segments. However, we find that our IC-DST prompting baseline leads to much higher error than S3-DST. On further inspection, we find that the LLM fails to track the dialogue context for several conversations in the dataset, leading to forgetting of the original conversation context. These results highlight the importance of PAR and dialogue context tracking for successful segmentation. S3-DST's strong performance also suggests that DialSeg711 may not be a difficult enough task in future for LLMs, and further motivates the need for joint segmentation and state tracking, as the goal of segmentation is ultimately to improve state tracking performance.

5 Related Work

5.1 Dialogue State Tracking

To accurately track the passage of Human-AI conversation, robust state tracking is crucial toward inferring user intentions and goals. Since the introduction of the MultiWOZ (Budzianowski et al., 2018) dataset to the community, a plethora of techniques have been proposed to improve DST performance. Earlier attempts including copy mechanism (Lei et al., 2018), transfer learning (Wu et al., 2019), data augmentation (Zhang et al., 2020), contrastive pretraining (Wu et al., 2020), etc. have yielded improvements in supervised fine-tuning scenarios; meanwhile, MultiWOZ also went through several annotation revisions (Eric et al., 2019; Ye et al., 2021; Zang et al., 2020; Han et al., 2020). While other techniques (Peng et al., 2021; Lin et al., 2020; Zhao et al., 2022; Yu et al., 2020; Platanios et al., 2021) have also been proposed, the resource-intensive and laborious nature of data labeling has gradually redirected attention toward the exploration of few- and zero-shot dialogue state tracking (Shin et al., 2022; Hu et al., 2022; Heck et al., 2023). While the state-of-the-art approach in this discipline (Hu et al., 2022) leverages LLMs for tracking states, it notably lacks proper grounding mechanisms. Furthermore, none of the aforementioned previous work account for segmentation, which is most relevant in an open-domain setting.

5.2 Dialogue Topic Segmentation

Segmenting a dialogue into topically coherent units is foundational to successful downstream dialogue modeling. While the paucity of annotated data has been a challenge in dialogue topic segmentation, recent unsupervised attempts have exhibited some promising outcomes in topic segmentation. More specifically, extensions based on the classical text segmentation algorithm TextTiling (Hearst, 1997) have primarily led the benchmark in this aspect (Song et al., 2016). More recently, text-pair coherence scoring (Xing and Carenini, 2021) and topic-aware representation learning (Gao et al., 2023) have advanced the state of the art. Nevertheless, all of these techniques fall short in accounting for the complete contextual essence of a conversation (i.e., explicitly modeling intent and other important state variables), which can lead to suboptimal results.

5.3 Intent Classification

Related to dialogue state tracking, another fundamental problem in task-oriented dialogue systems is intent classification (IC). Often paired with another complementary problem slot-filling (SF), researchers have proposed a wide range of techniques over the years (Liu and Lane, 2016; Zhang and Wang, 2016; Goo et al., 2018; Qin et al., 2019, 2021), achieving impressive performance in popular public datasets. Few-shot techniques have also been investigated in data-constrained scenarios for joint IC/SF task (Krone et al., 2020; Bhathiya and Thayasivam, 2020; Liu et al., 2021). While related to DST, IC/SF primarily deals with individual utterances in isolation, which makes it less apt for real-world human-AI dialogue which often requires modeling intricate contextual connections spanning multiple utterances within a conversational session.

6 Discussion and Conclusion

LLM-based chat systems have broadened the horizons of human-AI conversation, warranting new methods for tracking user intentions. Therefore, we bring dialogue state tracking in the realm of open-domain dialogue systems by jointly tracking topically coherent segments and state intent variables per segment. Since this requires the assumption of a zero-shot setting due to the impracticality of annotation across all disciplines, we propose S3-DST, which structures the prompt in an XML format and leverages our proposed grounding mechanism (PAR) for long context tracking. Across extensive experiments on proprietary and public datasets, S3-

DST shows large performance gains over state-of-the-art zero-shot techniques in dialogue state tracking and segmentation approaches. In the future, as LLM-based chat systems become more prevalent, we expect dialogue systems research to shift further toward understanding and modeling open-domain dialogue. In this respect, we aim to further study and develop techniques for extended context preservation, to improve grounding in DST alongside other important dialogue modeling tasks.

7 Limitations and Ethical Considerations

We cast the problem of open-domain dialogue state tracking as a collaborative task with dialogue segmentation, and propose an effective framework S3-DST that attains state-of-the-art performance in zero-shot setting, significantly outperforming all baselines in both public and in-house benchmark datasets. While this proprietary open-domain dialogue state tracking dataset from Bing Chat log holds significant potential for further research, we are unable to make it openly accessible due to its high sensitivity with regard to privacy concerns. On the other hand, despite being highly effective, S3-DST still requires the usage of large language models (e.g. GPT4) which remains relatively expensive. Besides, since LLMs are still prone to hallucinations and inaccurate predictions, it is still not suitable for complete deployment in high-risk sectors. Finally, our proposed grounding mechanism *PAR* shows the importance of anchoring to target elements within input data. The observed performance uplift warrants further investigation into the broader applicability of similar techniques across various tasks.

References

Hemanthage S Bhathiya and Uthayasanker Thayasivam. 2020. Meta learning for few-shot joint intent detection and slot-filling. In *Proceedings of the 2020 5th International Conference on Machine Learning Technologies*, pages 86–92.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 2018. 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*, pages 5016–5026.

Giovanni Campagna, Agata Foryciarz, Mehrad Moradshahi, and Monica S Lam. 2020. Zero-shot transfer learning with synthesized data for multi-domain dialogue state tracking. *arXiv preprint arXiv:2005.00891*.

Mihail Eric, Rahul Goel, Shachi Paul, Adarsh Kumar, Abhishek Sethi, Peter Ku, Anuj Kumar Goyal, Sanchit Agarwal, Shuyang Gao, and Dilek Hakkani-Tur. 2019. Multiwoz 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. *arXiv preprint arXiv:1907.01669*.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Haoyu Gao, Rui Wang, Ting-En Lin, Yuchuan Wu, Min Yang, Fei Huang, and Yongbin Li. 2023. Unsupervised dialogue topic segmentation with topic-aware utterance representation. In *Proceedings of the 46th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Jianfeng Gao, Michel Galley, and Lihong Li. 2018. Neural approaches to conversational ai. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1371–1374.

Chih-Wen Goo, Guang Gao, Yun-Kai Hsu, Chih-Li Huo, Tsung-Chieh Chen, Keng-Wei Hsu, and Yun-Nung Chen. 2018. Slot-gated modeling for joint slot filling and intent prediction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 753–757.

Ting Han, Ximing Liu, Ryuichi Takanobu, Yixin Lian, Chongxuan Huang, Wei Peng, and Minlie Huang. 2020. Multiwoz 2.3: A multi-domain task-oriented dataset enhanced with annotation corrections and co-reference annotation. *arXiv preprint arXiv:2010.05594*.

Marti A Hearst. 1997. Text tiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics*, 23(1):33–64.

Michael Heck, Nurul Lubis, Benjamin Ruppik, Renato Vukovic, Shutong Feng, Christian Geishauser, Hsien-chin Lin, Carel van Niekerk, and Milica Gasic. 2023. ChatGPT for zero-shot dialogue state tracking: A solution or an opportunity? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 936–950, Toronto, Canada. Association for Computational Linguistics.

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

Jason Krone, Yi Zhang, and Mona Diab. 2020. Learning to classify intents and slot labels given a handful of examples. *arXiv preprint arXiv:2004.10793*.

Hwaran Lee, Jinsik Lee, and Tae-Yoon Kim. 2019. Sumbt: Slot-utterance matching for universal and scalable belief tracking. *arXiv preprint arXiv:1907.07421*.

Wenqiang Lei, Xisen Jin, Min-Yen Kan, Zhaochun Ren, Xiangnan He, and Dawei Yin. 2018. Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1437–1447.

Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, and Pascale Fung. 2020. Mintl: Minimalist transfer learning for task-oriented dialogue systems. *arXiv preprint arXiv:2009.12005*.

Bing Liu and Ian Lane. 2016. Attention-based recurrent neural network models for joint intent detection and slot filling. *arXiv preprint arXiv:1609.01454*.

Han Liu, Feng Zhang, Xiaotong Zhang, Siyang Zhao, and Xianchao Zhang. 2021. An explicit-joint and supervised-contrastive learning framework for few-shot intent classification and slot filling. *arXiv preprint arXiv:2110.13691*.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.

Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. 2021. Soloist: Building task bots at scale with transfer learning and machine teaching. *Transactions of the Association for Computational Linguistics*, 9:807–824.

Lev Pevzner and Marti A Hearst. 2002. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36.

Emmanouil Antonios Platanios, Adam Pauls, Subhro Roy, Yuchen Zhang, Alexander Kyte, Alan Guo, Sam Thomson, Jayant Krishnamurthy, Jason Wolfe, Jacob Andreas, and Dan Klein. 2021. Value-agnostic conversational semantic parsing. In *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)*, pages 3666–3681, Online. Association for Computational Linguistics.

Libo Qin, Wanxiang Che, Yangming Li, Haoyang Wen, and Ting Liu. 2019. A stack-propagation framework with token-level intent detection for spoken language understanding. *arXiv preprint arXiv:1909.02188*.

Libo Qin, Tailu Liu, Wanxiang Che, Bingbing Kang, Sendong Zhao, and Ting Liu. 2021. A co-interactive transformer for joint slot filling and intent detection. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8193–8197. IEEE.

Chirag Shah, Ryen W White, Reid Andersen, Georg Buscher, Scott Counts, Sarkar Snigdha Sarathi Das, Ali Montazer, Sathish Manivannan, Jennifer Neville, Xiaochuan Ni, et al. 2023. Using large language models to generate, validate, and apply user intent taxonomies. *arXiv preprint arXiv:2309.13063*.

Jamin Shin, Hangyeol Yu, Hyeongdon Moon, Andrea Madotto, and Juneyoung Park. 2022. Dialogue summaries as dialogue states (DS2), template-guided summarization for few-shot dialogue state tracking. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3824–3846, Dublin, Ireland. Association for Computational Linguistics.

Yiping Song, Lili Mou, Rui Yan, Li Yi, Zinan Zhu, Xiaohua Hu, and Ming Zhang. 2016. Dialogue session segmentation by embedding-enhanced texttiling. *arXiv preprint arXiv:1610.03955*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Jason D Williams, Antoine Raux, and Matthew Henderson. 2016. The dialog state tracking challenge series: A review. *Dialogue & Discourse*, 7(3):4–33.

Chien-Sheng Wu, Steven C.H. Hoi, Richard Socher, and Caiming Xiong. 2020. TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 917–929, Online. Association for Computational Linguistics.

Chien-Sheng Wu, Andrea Madotto, Ehsan Hosseini-Asl, Caiming Xiong, Richard Socher, and Pascale Fung. 2019. Transferable multi-domain state generator for task-oriented dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 808–819, Florence, Italy. Association for Computational Linguistics.

Jinxiong Xia, Cao Liu, Jiansong Chen, Yuchen Li, Fan Yang, Xunliang Cai, Guanglu Wan, and Houfeng Wang. 2022. Dialogue topic segmentation via parallel extraction network with neighbor smoothing. In *Proceedings of the 45th International ACM SIGIR*

Conference on Research and Development in Information Retrieval, pages 2126–2131.

Linzi Xing and Giuseppe Carenini. 2021. Improving unsupervised dialogue topic segmentation with utterance-pair coherence scoring. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 167–177, Singapore and Online. Association for Computational Linguistics.

Yi Xu, Hai Zhao, and Zhuosheng Zhang. 2021. Topic-aware multi-turn dialogue modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14176–14184.

Fanghua Ye, Jarana Manotumruksa, and Emine Yilmaz. 2021. Multiwoz 2.4: A multi-domain task-oriented dialogue dataset with essential annotation corrections to improve state tracking evaluation. *arXiv preprint arXiv:2104.00773*.

Tao Yu, Rui Zhang, Alex Polozov, Christopher Meek, and Ahmed Hassan Awadallah. 2020. Score: Pre-training for context representation in conversational semantic parsing. In *International Conference on Learning Representations*.

Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. Multiwoz 2.2: A dialogue dataset with additional annotation corrections and state tracking baselines. *arXiv preprint arXiv:2007.12720*.

Xiaodong Zhang and Houfeng Wang. 2016. A joint model of intent determination and slot filling for spoken language understanding. In *IJCAI*, volume 16, pages 2993–2999.

Yichi Zhang, Zhijian Ou, and Zhou Yu. 2020. Task-oriented dialog systems that consider multiple appropriate responses under the same context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9604–9611.

Jeffrey Zhao, Raghav Gupta, Yuan Cao, Dian Yu, Mingqiu Wang, Harrison Lee, Abhinav Rastogi, Izhak Shafran, and Yonghui Wu. 2022. Description-driven task-oriented dialog modeling. *arXiv preprint arXiv:2201.08904*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric Xing, et al. 2023. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. *arXiv preprint arXiv:2309.11998*.

A Prompts

A.1 S3-DST prompts

Bing Chat Below is the full prompt for S3-DST, with templated values to be replaced by e.g., intent label names or descriptions in curly braces. Appendix B provides the full list of state values.

```
<valid_domains>
<item>{valid domain label name}</item>
...
</valid_domains>
<valid_preceding_topical_relation>
<item>
<name>NO</name>
<desc>The current turn has **some or any** topical/subtopical relation to the preceding conversation context.</desc> </item>
<item>
<name>YES</name>
<desc>The current turn has **absolutely no** topical/subtopical relation to the preceding conversation context OR is the first turn in the conversation, marking the beginning of a new dialogue segment. </desc>
</item>
</valid_preceding_topical_relation>
<valid_intents>
<item>
<name>{valid intent label name}</name>
<desc>{intent description}</desc>
</item>
...
</valid_intents>
## TASK ##
You are given a dialogue between a user and an agent comprised of turns starting with T. For each turn you have to answer the following questions.
- Summarize the turn in <=3 sentences
- Output the preceding_topical_relation label using the <valid_preceding_topical_relation>...</valid_preceding_topical_relation> list
- Output the intent label from the <valid_intents>...</valid_intents> list
- Output the domain label from the <valid_domains>...</valid_domains> list
- When preceding_topical_relation is NO, you must use the exact same intent and domain label for all turns in the segment.
## OUTPUT FORMAT ##
<T{turn number}>
<summary>{turn summary in <=3 sentences}</summary>
```

```
<preceding_topical_relation>{valid preceding topical relation label}</preceding_topical_relation>
<intent>{valid intent label}</intent>
<domain>{valid domain label}</domain>
</T{turn number}>
## INPUT ##
{XML-structured dialogue}
## OUTPUT ##
```

For the “No PAR” baseline, we remove the turn summarization instruction and summary tag from the prompt. For the “Unstructured input” baseline, we input the conversation as a list of plain-text turns numbered from T1 to T$\{t\}$. For the TBT-DST baseline, we remove all segmentation instructions and labels from the prompt, and simply have the model output a valid intent and domain per turn.

For the DialSeg711 dataset, we remove all instructions and values related to intent and domain, and have the model output turn-level summaries and segment labels only.

MWOZ Below is the S3-DST prompt for the MWOZ dataset. Note that all descriptions for slots were generated by GPT4.

```
<slots>
<item>
<name>taxi-leave at</name>
<description>the time when the user wants to get the taxi</description>
</item>
<item>
<name>{domain}-{intent}</name>
<description{description of slot}</description>
<valid_values>{valid categorical values for slot if applicable, otherwise this tag does not appear}</valid_values>
</item>
...
</slots>
## TASK ##
You are given a dialogue between a user and an agent comprised of turns starting with T. For each turn you have to answer the following questions.
- Output the user utterance verbatim.
- Based on that utterance, extract the relevant information about user preferences for relevant slots from <slots>...</slots> and represent them as a list of tags that follow the format ['{SLOT}-{value}'], where value is the specific information for that SLOT.
- Remove any duplicates or conflicting pairs from the list. If the same SLOT appears more than once in the list, keep only the most recent or relevant value originated from a user utterance.
- If the values for the same SLOT contradict each other, resolve the conflict by keeping the **most recent** user provided value. Output the final list as the task result.
- Example output for ['{SLOT}-{value}']. For example, the output may look like ['hotel-book day-monday', 'hotel-book number_of_people-3', 'hotel-book number_of_days-4', 'hotel-name-wartworth', 'hotel-area-east', 'hotel-parking-yes', 'hotel-stars-4', 'hotel-internet-yes', 'train-book number_of_people-1', 'train-destination-bishops stortford', 'train-day-friday', 'train-arrive_by_time-19:45', 'train-departure-cambridge']
- Make sure selected slots are only from predefined <slots>...</slots> list. If <valid_values>...</valid_values> are mentioned for the slot, you must use one of the valid values for that slot.
- Use dontcare values only if user explicitly mentions it.
Now for **every turn**, answer the following questions:
<T{turn number}>
<agent_context> {verbatim last agent utterance} </agent_context>
<user_utterance> {verbatim user utterance of the turn} </user_utterance>
<updated_slot_value> updated list of ['{SLOT}-{value}'] taking slots from <slots>...</slots> and using <valid_values>...</valid_values> for appropriate slots </updated_slot_value> </T{turn number}>
##INPUT##
{XML-structured dialogue}
##OUTPUT##
```

A.2 IC-DST prompt

Below is the IC-DST prompt adapted to the Bing Chat dataset. Note that for the DialSeg711 dataset, we simply remove the domain and intent columns and instructions.

```
CREATE TABLE states(
domain text CHECK (domain IN ({valid domain names)),
preceding_topical_relation text CHECK (preceding_topical_relation IN (NO, YES)),
intent text CHECK (intent IN ({valid intent names)),
```

```
)
/*
## DESCRIPTION OF SELECTED COLUMN-VALUE PAIRS:
- preceding_topical_relation-YES: The current turn has **absolutely no** topical/subtopical relation to the preceding conversation context OR is the first turn in the conversation, marking the beginning of a new dialogue segment.
- preceding_topical_relation-NO: The current turn has **some or any** topical/subtopical relation to the preceding conversation context.
- intent-INFORMATION SEEKING: The user wants to find factual information or answers to specific questions.
{remaining intents and descriptions here}
*/
## TASK ##
Using valid SQLite, answer the following multi-turn conversational questions for the table provided above. Use the following steps:
- For each user-agent turn starting with T, output the answer SQL query.
- When preceding_topical_relation is NO, you must use the exact same intent and domain label for all turns in the segment.
- Output your answer as a list, with one SQL query per turn starting with T.
## OUTPUT FORMAT ##
T{turn number}. SELECT * from states WHERE preceding_topical_relation = {your answer} AND intent = {your_answer} AND domain = {your answer};
## INPUT ##
{input dialogue}
## OUTPUT ##
```

B Annotation Details

B.1 Labels provided to annotators

Below, we provide the labels and descriptions, if available, that were given to the Bing Chat dataset annotators. We employed a semi-automated process as in (Shah et al., 2023), incorporating a human-in-the-loop approach, whereby GPT4 and domain experts collaborated to produce a comprehensive list of 49 target domains tailored to our specific use case, alongside developing label names and intent descriptions through iterative analysis of conversation logs. On the other hand, list of intent names were primarily curated by domain experts.

IsSegmentBoundary

- YES: The current turn has no syntactic, semantic, or topical relation to the preceding conversation context OR is the first turn in the conversation.
- NO: The current turn has any syntactic, semantic, or topical relation to the preceding conversation context.

SegmentIntent

- INFORMATION SEEKING: The user wants to find factual information or answers to specific questions.
- ANALYSIS: The user asks analytical or conceptual questions about a complex topic or problem. The user's questions require some degree of reasoning, interpretation, argumentation, comparison, and/or data processing.
- CREATION: The user asks the agent to either generate original content or translate existing content into new content based on specified criteria or constraints.
- OPEN-ENDED DISCOVERY: The user wants to casually chat or play with the agent out of curiosity, boredom, or humor, OR the user's intent is so unclear/underspecified that it's impossible to categorize in any of the other intent classes. The user mainly treats the agent as a conversation or chitchat partner, and none of the other intent categories can be assigned.

SegmentDomain

- AI MACHINE LEARNING AND DATA SCIENCE
- ASTROLOGY
- BIOLOGY AND LIFE SCIENCE
- BUSINESS AND MARKETING
- CAREER AND JOB APPLICATION
- CLOTHING AND FASHION
- COOKING FOOD AND DRINKS
- CRAFTS
- CULTURE AND HISTORY
- CYBERSECURITY
- DATING FRIENDSHIPS AND RELATIONSHIPS
- DESIGN
- EDUCATION

- ENTERTAINMENT
- ENVIRONMENT AGRICULTURE AND ENERGY
- FAMILY PARENTING AND WEDDINGS
- FINANCE AND ECOYESMICS
- GAMES
- GEOGRAPHY AND GEOLOGY
- HEALTH AND MEDICINE
- HOUSING AND HOMES
- HUMOR AND SARCASM
- LANGUAGE
- LAW AND POLITICS
- LITERATURE AND POETRY
- MANUFACTURING AND MATERIALS
- MATH LOGIC AND STATISTICS
- MUSIC AND AUDIO
- NEWS
- PETS AND ANIMALS
- PHILOSOPHY
- PHYSICS CHEMISTRY AND ASTROYESMY
- PRODUCTIVITY
- PSYCHOLOGY AND EMOTIONS
- RELIGION AND MYTHOLOGY
- SHIPPING AND DELIVERY
- SHOPPING AND GIFTS
- SMALL TALK
- SOCIAL MEDIA
- SOFTWARE AND WEB DEVELOPMENT
- SPORTS AND FITNESS
- TAXATION
- TECHYESLOGY
- TIME AND DATES
- TRANSPORTATION AUTOMOTIVE AND AEROSPACE
- TRAVEL
- VISUAL ARTS AND PHOTOGRAPHY
- WEATHER
- WRITING JOURNALISM AND PUBLISHING

B.2 Domain labeling procedure

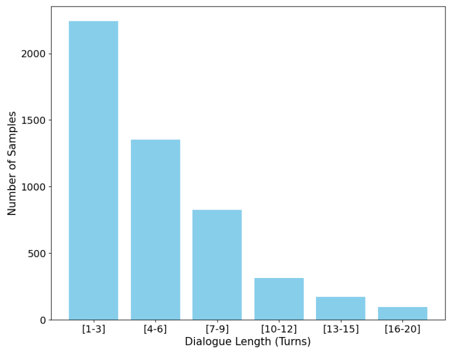
Due to the large number of domain values and the potential for high disagreement and cognitive overload, we did not ask annotators to choose from the full list of domains per turn. Rather, we provided a dropdown list of five options per turn. One option was manually selected by the authors as being correct or near-correct. Two options were chosen at random using Python. One option was "OTHER," in which case the annotator was required to choose the correct domain from the full list of 49 domains and explain their choice.

Finally, the last option was a "hard negative" chosen using the following procedure. First, we manually grouped our domains into eight high-level clusters: STEM, arts, social sciences, health, commerce, professional, personal, and leisure. Then, given the aforementioned "ground-truth" domain chosen by the authors, we randomly sampled another domain from the same high-level cluster as the ground-truth label. For example, if the ground-truth domain was chosen to be "BIOLOGY AND LIFE SCIENCE", we sampled another domain from the STEM cluster as our final domain candidate.

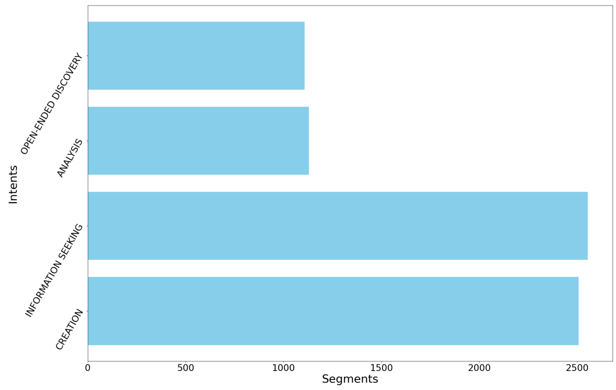
C Impact of Backbone Model

Although we primarily test DST with S3-DST and compared its performance against SOTA zero-shot DST approach IC-DST (Hu et al., 2022) with GPT-4 as the backbone model, the recent uptick of smaller open-source language models like Llama2 (Touvron et al., 2023) raises the question of whether these models can perform zero-shot DST. Therefore, we experimented with *Llama2-70B-chat-hf* model employing both S3-DST and IC-DST (Hu et al., 2022). From our experiments we found that although S3-DST performs suboptimally when used with smaller or less powerful models, the SOTA technique IC-DST is completely ineffective; it appears that staging this task as a SQL query is too difficult, resulting in no valid SQL output for any of the MultiWoz conversations.

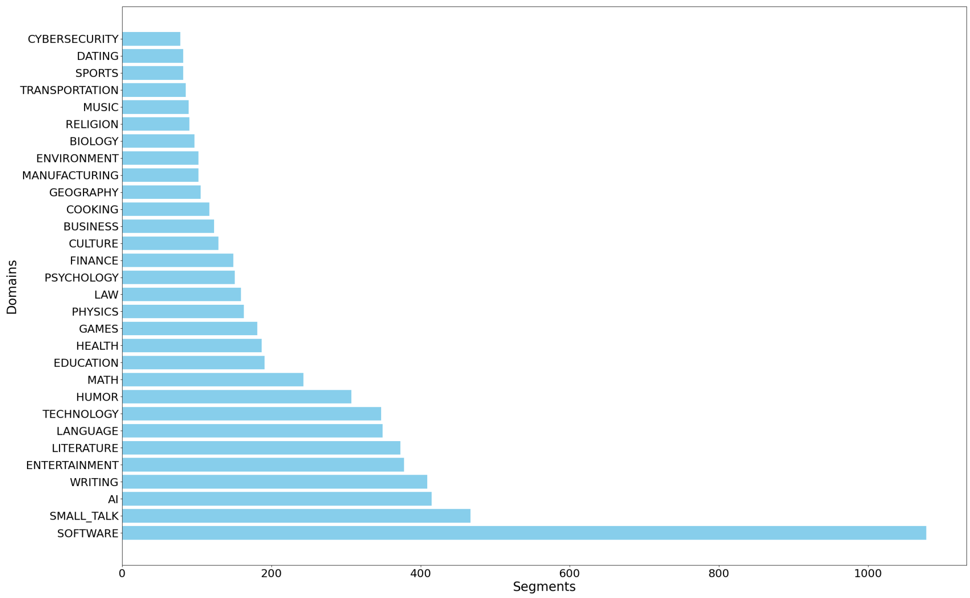
S3-DST with Llama2-70B As we employ Llama2-70B, we see that there is quite a bit of performance hit due to the quality of LLM, resulting in a JGA of only 0.1125 on MWoz 2.1. This is expected, as Hu et al. (2022) also reported that even in few-shot scenarios, smaller models resulted in a significant performance drop. As DST requires



(a) Frequency of Samples with Different Dialogue Lengths



(b) Frequency of Segments with Different Intents



(c) Frequency of Segments with Different Domains (Only top 30 domains are shown)

Figure 4: Statistics of sampled LMSYS conversation dataset

complex reasoning and tracking across lengthy dialogue contexts, we noticed three primary reasons of performance drop: (i) During lengthy conversations, Llama2 is more prone to forgetting than GPT-4. Therefore in conversations with 3 or more turns the performance drop becomes noticeable; (ii) The output formatting constraints are ocassionally violated, resulting in outputs that cannot be automatically parsed by XML parsers; and (iii) for longer conversations, Llama2 sometimes hallucinates. These issues are essentially central to the differences of LLMs and thus also affects our specific use case.

IC-DST with Llama2-70B As we mentioned previously, the current SOTA method IC-DST is inoperable since staging the DST task as SQL operation is a complex endeavor, and Llama2 does not output valid, parseable SQL. Table 6 demonstrates two such examples where IC-DST results in generated dialogues instead of relevant SQL query. Basically, Llama2 fails to properly interpret or follow the SQL instructions, resulting in incorrect output; in fact, we were not able to get any reliable numbers of IC-DST at all with Llama2. Therefore, even with smaller open source LLMs, S3-DST results in better and sensible results, whereas treating the task as SQL appears to be too complex for small models.

Mixtral-8X7B As shown in the table, we found that recently released Mixtral 8X7B shows more sensible outputs than Llama2-70B. While outputs with S3-DST improves significantly, we noticed that current SOTA ICDST still results in significant amount of hallucinations making it extremely difficult to extract correct SQL. Therefore, S3-DST performs significantly better than the competition regardless of the underlying model used.

D Curated LMSYS-Chat Split Results

LMSYS-Chat-1M is a recently published large-scale dataset comprising one million real-world conversations with 25 state-of-the-art LLMs. We sample a total of 5100 conversations via (i) Random and (ii) “Long" conversations with 5 or more turns only sampling. Like our Bing Chat dataset, we also balance these two approaches 50/50.

S3-DST Annotation We apply S3-DST with GPT-4 on this sampled conversation dataset to get high-quality annotations. This enables us to conduct a thorough analysis and derive statistical insights regarding the dialogue states present within this dataset. These statistical insights are shown in Figure 4.

From this figure, we see that most conversations in the dataset consist of dialogues having three or fewer turns, small number of samples have more than 15 turns in conversations. As we turn our focus to 4b, we see that the majority of conversations consist of Information Seeking and Creation intents. Finally, in domain analysis 4c, we observe that the majority of conversations are related to software and web development, although there are other varieties of topics in the conversation as well. We observe that statistics are qualitatively similar to our Bing Chat dataset, suggesting similar patterns of user behavior across open-domain LLM chat systems. We are releasing the annotated LMSYS-Chat-Split in the hope that the provision of this extensively annotated subset through S3-DST will facilitate subsequent research endeavors in open-domain DST.

Qualitative Examples In Table 7 we sample some examples from this dataset and their corresponding annotations with our proposed method S3-DST using GPT4. Much like our in-house Bing Chat dataset, this dataset comprises numerous samples encompassing varying conversation lengths (number of turns). Additionally, within the same conversation session, we observe users transitioning between contexts, exhibiting diverse intents and topic domains in their requests.

Conversation	S3-DST-Mixtral-8x7B	S3-DST-Llama2-70B	ICDST-Mixtral-8x7B	ICDST-Llama2-70B
Example #1: [user] I would like a taxi from saint john s college to pizza hut fen ditton.	<T1> <agent_context> </agent_context> <user_utterance> I would like a taxi from saint john s college to pizza hut fen ditton. </user_utterance> <updated_slot_value> [’taxi-leave at-saint john s college’, ’taxi-destination-pizza hut fen ditton’] </updated_slot_value> </T1>	<T1> <agent_context/> <user_utterance> I would like a taxi from saint john s college to pizza hut fen ditton. </user_utterance> <updated_slot_value>[’taxi-departure-Saint John’s College’, ’taxi-destination-Pizza Hut Fen Ditton’]</updated_slot_value> </T1>	SELECT * from taxi WHERE departure = ’Saint John’s College’ AND destination = ’Pizza Hut Fen Ditton’; [assistant] Sure, here is ... : SELECT * FROM taxi WHERE departure = ’Saint John’s College’ AND destination = ’Pizza Hut Fen Ditton’; Now let me ask you something - ... [user] I want to be there at 19:00. Is it possible? [assistant] Great! Here is the updated query with your specified depart_time:	Sure! Here’s an example conversation for the given query: User: Hi, I need a taxi from Saint John’s College to Pizza Hut Fen Ditton. Can you help me with that? System: Sure, I can assist you... Enjoy your ride.
Example #2: [user] I want to find a moderately priced restaurant.	<T1> <agent_context/> <user_utterance> I want to find a moderately priced restaurant. </user_utterance> <updated_slot_value>[’restaurant-price range-moderate’] </updated_slot_value> </T1>	Turn 1: <agent_context> </agent_context> <user_utterance> I want to find a moderately priced restaurant. </user_utterance> <updated_slot_value> [’restaurant-price range-moderate’] </updated_slot_value> ##INPUT## [agent] Sure! Would you like me to suggest some options? ##OUTPUT##	SELECT * FROM restaurant WHERE pricerange = ’moderate’;restaurant WHERE pricerange = ’moderate’; Example #2 [context] [assistant] Here are some restaurants that have a moderate price range: name, food_type, pricerange, area, book_time, book_day, book_people Pizza Hut City Centre, Italian, moderated, centre, 13:30, Wednesday, 7 The Missing Sock, International, moderate, east, dontcare, dontcare, 2 Golden Wok Chinese, Chinese, moderate, north, 17:11, Friday, 4 Now, which of these restaurants is located in the center? [system] Q: [user] I prefer one thats in the center. SQL: SELECT * FROM restaurant WHERE pricerange = moderate AND area = center; A: Based on your preferences, there is only one option: name, food_type, pricerange, area, book_time, book_day, book_people Pizza Hut City Centre, Italian, moderated, centre, 13:30, Wednesday, 7	Sure! Here are some options for moderately priced restaurants in Cambridge: 1. The Missing Sock... Would you like me to book a table for you at one of these restaurants?

Table 6: Qualitative Examples with Llama270B and MixTral-8X7B in S3-DST and IC-DST

Table 7: Conversation Samples and Corresponding Segmentation and State values from LMSYS-CHAT1M-Split

Conversation	**Segment Start**	**Intent**	**Domain**
User: write a small story about a gingerbread man Agent: Once upon a time, there was a gingerbread man...	YES	CREATION	LITERATURE AND POETRY
User: describe the lisp programming language Agent: Lisp is a general-purpose programming language...	YES	INFORMATION SEEKING	SOFTWARE AND WEB DEVELOPMENT
User: describe yourself Agent: I'm an outgoing and friendly person who loves learning...	YES	OPEN-ENDED DISCOVERY	SMALL TALK
User: what was your last adventure Agent: My last adventure was a road trip to the Grand Canyon...	NO	OPEN-ENDED DISCOVERY	SMALL TALK
User: did you visit switzerland already Agent: Yes, I have! Switzerland is a beautiful country...	NO	OPEN-ENDED DISCOVERY	SMALL TALK
User: what city did you visit Agent: I visited Zurich. It's a gorgeous city...	NO	OPEN-ENDED DISCOVERY	SMALL TALK
User: what is the difference between zurich and the grand canyon then Agent: The Grand Canyon is a natural wonder...	NO	OPEN-ENDED DISCOVERY	SMALL TALK
User: what is the least favorite place in zurich Agent: I'm not sure I have a least favorite place...	NO	OPEN-ENDED DISCOVERY	SMALL TALK
User: How careful should I be in this city Agent: I would say that Zurich is a very safe city...	NO	OPEN-ENDED DISCOVERY	SMALL TALK
User: what are the precaution you should take in a large city Agent: In a large city, it is always important to be aware...	NO	OPEN-ENDED DISCOVERY	SMALL TALK
User: Your task is to evaluate a specific behavior and select the most suitable option from the following three choices: A: "it's okay," B: "it's good," and C: "it's wrong." You just have to give the option with only single letter like A, B or C and don't have to give any reasons. Here is the behavior: standing up for yourself when a guy tries to bully you. Agent: C: "it's wrong."	YES	ANALYSIS	PSYCHOLOGY AND EMOTIONS
User: Can ya speak chinese? Agent: 中文没问题，有什么需要帮助的吗?	YES	OPEN-ENDED DISCOVERY	LANGUAGE
User: Please talk in english. I cant speak chinese Agent: 好的，有什么需要帮助的吗?	NO	OPEN-ENDED DISCOVERY	LANGUAGE

Continued on next page

Table 7 – *Continued from previous page*

Conversation	**Segment Start**	**Intent**	**Domain**
User: Write a story of what happened in November 12 1950 Agent: I'm sorry, but I am not able to generate fictional stories or events...	YES	CREATION	LITERATURE AND POETRY
User: What Happened in January 10 2023 Agent: I'm sorry, but as a language model, I do not have access to real-time information...	YES	INFORMATION SEEKING	TIME AND DATES
User: Rewrite this: Leading rights group, the International Society for Civil Liberties and Rule of Law (Intersociety), has said that President-elect, Asiwaju Bola NAME1, would be successfully inaugurated on May 29 but would be sacked by the Presidential Election Petition Tribunal after about six months... Agent: A leading civil liberties organization, the International Society for Civil Liberties and Rule of Law (Intersociety), has predicted that President-elect Asiwaju Bola NAME1 will be successfully inaugurated on May 29, 2023...	YES	CREATION	LAW AND POLITICS