NATCS: Eliciting Natural Customer Support Dialogues

James Gung, Emily Moeng, Wesley Rose
Arshit Gupta, Yi Zhang, and Saab Mansour
AWS AI Labs
{gungj,emimoeng,rosewes,arshig,yizhngn,saabm}@amazon.com

Abstract

Despite growing interest in applications based on natural customer support conversations, there exist remarkably few publicly available datasets that reflect the expected characteristics of conversations in these settings. Existing task-oriented dialogue datasets, which were collected to benchmark dialogue systems mainly in written human-to-bot settings, are not representative of real customer support conversations and do not provide realistic benchmarks for systems that are applied to natural data. To address this gap, we introduce NATCS, a multi-domain collection of spoken customer service conversations. We describe our process for collecting synthetic conversations between customers and agents based on natural language phenomena observed in real conversations. Compared to previous dialogue datasets, the conversations collected with our approach are more representative of real human-to-human conversations along multiple metrics. Finally, we demonstrate potential uses of NATCS, including dialogue act classification and intent induction from conversations as potential applications, showing that dialogue act annotations in NATCS provide more effective training data for modeling real conversations compared to existing synthetic written datasets. We publicly release NATCS to facilitate research in natural dialog systems[1].

1 Introduction

Applications that are applied to human-to-human customer support conversations have become increasingly popular in recent years. For example, assistive tools that aim to support human agents, provide analytics, and automate mundane tasks have become ubiquitous in industry applications (Amazon Contact Lens, 2023; Google Contact Center AI, 2023; Microsoft Digital Contact Center Platform, 2023). Despite this growing interest in data processing for natural customer service conversations, to the best of our knowledge, there exist no public datasets to facilitate open research in this area.

Existing dialogue datasets focusing on development and evaluation of task-oriented dialogue (TOD) systems contain conversations that are representative of human-to-bot (H2B) conversations adhering to restricted domains and schemas (Budzianowski et al., 2018; Rastogi et al., 2020). Realistic live conversations are difficult to simulate due to the training required to convincingly play the role of an expert customer support agent in non-trivial domains (Chen et al., 2021). Existing datasets are also primarily written rather than spoken conversations as this modality is cheaper to simultaneously collect and annotate asynchronously through crowdsourcing platforms.

To address these gaps, we present NATCS, a multi-domain dataset containing English conversations that simulate natural, human-to-human (H2H), two-party customer service interactions. First, we describe a self-collection dataset, $\text{NATCS}_{\text{SELF}}$, where we use a strict set of instructions asking participants to write both sides of a conversation as if it had been spoken. Second, we present a spoken dataset, $\text{NATCS}_{\text{SPOKE}}$, in which pairs of participants were each given detailed instructions and asked to carry out and record conversations which were subsequently transcribed. We observe that the resulting conversations in NATCS share more characteristics with real customer service conversations than pre-existing dialogue datasets in terms of diversity and modeling difficulty.

We annotate a subset of the conversations in two ways: (1) we collect task-oriented dialogue act annotations, which label utterances that are important for moving the customer's goal forward and (2) we categorize and label customer goals and goal-related information with an open intent and slot

[1] https://github.com/amazon-science/dstc11-track2-intent-induction.

Findings of the Association for Computational Linguistics: ACL 2023, pages 9652–9677
July 9-14, 2023

schema, mimicking the process for building a TOD system based on natural conversations. We find that classifiers trained with the resulting dialogue act annotations have improvements in accuracy on real data as compared to models trained with pre-existing TOD data.

Our main contributions are threefold:

- We present NATCS, a multi-domain dialogue dataset containing conversations that mimic spoken H2H customer service interactions, addressing a major gap in existing datasets.
- We show that NATCS is more representative of conversations from real customer support centers than pre-existing dialogue datasets by measuring multiple characteristics related to realism, diversity and modeling difficulty.
- We provide TOD dialogue act and intent/slot annotations on a subset of the conversations to facilitate evaluation and development of systems that aim to learn from real conversations (such as intent and slot induction), empirically demonstrating the efficacy of the dialogue annotations on real data.

Our paper is structured as follows: In Section 2, we review other dataset collection methods and approaches for evaluating dialogue quality. In Section 3, we describe how NATCS conversations and annotations were collected. In Section 4, we compare NATCS with pre-existing dialogue datasets as well as real H2H customer service conversations. Finally, in Section 5, we further motivate the dataset through two potential downstream applications, task-oriented dialogue act classification and intent induction evaluation.

2 Related Work

Dialogue Dataset Collection The goal of NATCS, to produce conversations that emulate real spoken customer service interactions, differs substantially from previous synthetic dialogue dataset collections. Previous synthetic, goal-oriented datasets have used the *Wizard of Oz* framework (Kelley, 1984). This framework calls for one person to interact with what they think is a computer, but is actually controlled by another person, thus encouraging a human-to-bot (H2b) style. Wen et al. (2016) define a specific version of this approach to be used with crowdsourced workers to produce synthetic, task-oriented datasets. This approach has since been adopted as a standard method to collect goal-oriented conversations (Peskov et al., 2019; Byrne et al., 2019; Budzianowski et al., 2018; El Asri et al., 2017; Eric et al., 2017).

MultiDoGO (Peskov et al., 2019) (MDGO), SGD (Rastogi et al., 2020), and TaskMaster (Byrne et al., 2019) are particularly relevant comparisons to our work. MDGO explicitly encourages dialogue complexities, which serve as inspiration for our complexity-driven methodologies. SGD is interesting partially because of the scale of their collection, relevance of their task-oriented dialogue act and intent/slot annotations, and because they diverge from the common practice of using the Wizard-of-Oz methodology through the use of dialogue templates and paraphrasing. TaskMaster presents a methodology for self-collected dialogues, which is analogous to our $\text{NATCS}_{\text{SELF}}$ collection.

Despite these similarities, the methodology for NATCS differs significantly from any of these pre-existing datasets, because of both the target modality (spoken, or spoken-like conversations), and the setting (H2H instead of H2B).

Analysis of Dialogue Quality An important component of our work is the comparison of synthetic dialogue datasets with real data. We adopt multiple metrics for comparing NATCS with both real and previous TOD datasets. Byrne et al. (2019) use perplexity and BLEU score as stand-ins for "naturalness", with the logic that a dataset should be harder for a model to learn if it is more realistic, because realistic data tends to be more diverse than synthetic data. Previous collections of dialogue datasets also have made comparisons based on surface statistics such as the number of dialogues, turns, and unique tokens. Casanueva et al. (2022) compares intent classification datasets based on both lexical diversity metrics and a semantic diversity metric computed using a sentence embedding model.

There are a number of metrics more common in other sub-domains that may be useful for measuring naturalness in dialogues. The measure for textual lexical diversity (MTLD), discussed in depth by McCarthy and Jarvis (2010), provides a lexical diversity measure less biased by document length. Liao et al. (2017) introduce a dialog complexity metric, intended for analyzing real customer service conversations, which is computed by assigning importance measures to different terms, computing

Text	Annotations
A: Thank you for calling Intellibank. What can I do for you today?	ElicitIntent
C: Hi, I was trying to figure out how to create an account online, but I'm having some trouble.	InformIntent (SetUpOnlineBanking)
A: I'm sorry to hear that. I'd be happy to help you create an online account.	
C: Well, actually, the reason I called is to check the balance on my account. Could you help me with that?	InformIntent (CheckAccountBalance)
A: Oh, I see. Sure. Let me pull up your account so I can check your balance. May I have your account number$_{\text{AccountNumber}}$, please?	ElicitSlot
C: Sure. It's two one two three four five$_{\text{AccountNumber}}$.	InformSlot
A: Thanks. And is this account a checking$_{\text{TypeOfAccount}}$ or savings$_{\text{TypeOfAccount}}$?	ConfirmSlot, ElicitSlot
C: Oh, it's checking$_{\text{TypeOfAccount}}$.	InformSlot

Table 1: Example conversation from NATCS Banking. Conversations are annotated with dialogue act annotations such as InformIntent and ElicitIntent, intents such as SetUpOnlineBanking, and slots such as AccountNumber.

utterance-level complexity, and weighting the contribution of utterances based on their dialogue act tags. Hewitt and Beaver (2020) performed a thorough comparison of the style of human-to-human vs. human-to-bot conversations, using lexical diversity measures, syntactic complexity, and other dimensions like gratitude, sentiment and amount of profanity, though the data was not released.

3 Methodology

3.1 Collection Methods

We propose two collection methods as part of NATCS. We have three goals for the resulting conversations: (1) They should exhibit the spoken modality, (2) all conversations from each domain should seem to be from the same company, and (3) they should appear to be real, human-to-human conversations between a customer and an agent. We explore two methodologies, resulting in the NATCS$_{\text{SPOKE}}$ and NATCS$_{\text{SELF}}$ datasets, to weigh collection cost and complexity compared to dataset effectiveness.

To support goal (3), we propose a set of **discourse complexity** types. The motivation for providing specific discourse complexities is to encourage some of the noise and non-linearity present in real human-to-human conversations. Based on manual inspection of 10 transcribed conversations from a single commercial call center dataset, we identify a combination of human expressions (social niceties, emotionally-charged utterances), phenomena mimicking imperfect, non-linear thought processes (change of mind, forgetfulness/unknown terminology, unplanned conversational flows), reflections of the wider context surrounding the conversation (continuing from a previous conversation, pausing to find information), distinctions between speakers' knowledge bases, and the use of multiple requests in single utterances (stating multiple intents, providing multiple slot values). A list of these complexities along with estimated target percentages (minimum percent of conversations where these phenomena should be present) is provided in Figure 1. Descriptions and examples are provided in Appendix Table 15.

ChitChat (60%), FollowUpQuestion (30%), ImplicitDescriptiveIntent (30%), PauseForLookup (30%), BackgroundDetail (25%), MultiElicit (25%), SlotCorrection (20%), Overfill (15%), IntentChange (10%), MultiIntent (10%), MultiIntentUpfront (10%), MultiValue (10%), SlotChange (10%), Callback (5%), Frustration (5%), MissingInfoCantFulfill (5%), MissingInfoWorkaround (5%), SlotLookup (5%)

Figure 1: Discourse complexities and target percentages of conversations containing them used to encourage phenomena observed in real conversations. Percentages were estimated based on manual inspection of a small set of real conversations.

To achieve cross-dataset consistency, supporting goal (2), collectors are provided with mock company profiles, including name of the mock company, as well as mock product or service names with associated prices. Collectors are also provided with a schema of intents and associated slots. Some

flexibility is allowed in the slot schema to reflect real world situations where customers may not have all requested information on hand. Examples of company profiles and intent schemas are provided in Appendix Tables 13 and 14. For each conversation, we sample a set of minimum discourse complexity types. For example, one conversation could be assigned the target complexities of ChitChat, FollowUpQuestion, MultiElicit, and SlotLookup. Scenarios eliciting each of these complexity types are generated and provided to the participants.

For NATCS$_{\text{SPOKE}}$, one participant plays the part of the customer service representative ("agent"), and one participant plays the part of the customer ("customer"). The participants are recorded as they play-act the scenarios described on their instruction sheets from the same room. These audio recordings are then transcribed and annotated for actual complexities.

Given the time, cost, and complexity involved for the creation of the NATCS$_{\text{SPOKE}}$ datasets, as an alternative approach, we apply the NATCS$_{\text{SELF}}$ method. For NATCS$_{\text{SELF}}$, participants write self-dialogues as if they were spoken out loud. This method has the benefit of (1) not needing to be transcribed, and (2) requiring only one participant to create each conversation and therefore not requiring scheduling to match participants together. However, these rely on an understanding by the participants of the distinction between spoken and written modality data. The NATCS$_{\text{SELF}}$ method follows a similar set-up as the NATCS$_{\text{SPOKE}}$ method, except in addition to being provided with a set of target discourse complexities, participants are also provided with a set of **spoken form complexities**. While discourse complexities target discourse-related phenomena, spoken form complexities consist of phenomena specifically observed in spoken form speech. For this complexity type, we include phenomena such as hesitations or fillers ('um'), rambling, spelling, and backchanneling ('uh huh go on'). A list of these complexities along with target percentages is provided in Figure 4, and further examples are provided in Appendix Table 16.

3.2 Annotations

One goal of collecting realistic dialogues is to facilitate the development and evaluation of tools for building task-oriented dialogue systems from H2H conversations. To this end, we perform two types of annotations on a subset of NATCS: Dialogue Act (DA) annotations and Intent Classification and Slot Labeling (IC/SL) annotations.

IC/SL annotations are intended to label intents and slots, two key elements of many TOD systems. An intent is broadly a customer goal, and a slot is a smaller piece of information related to that goal. We use an open labelset, asking annotators to come up with specific labels for each intent and slot, such as "BookFlight" and "PreferredAirline" as opposed to simply "Intent" and "Slot". Annotators are instructed to label the same intent no more than once per conversation. For slots, we use the principle of labeling the smallest complete grammatical constituent that communicates the necessary information.

Our DA annotations are intended to identify utterances that move the dialog towards the customer's goal. TOD systems often support only a small set of dialogue acts that capture supported user and agent actions. For the agent, these may include eliciting the user's intent or asking for slot values associated with that intent (*ElicitIntent* and *ElicitSlot* respectively). For the user, such acts may include informing the agent of a new intent or providing relevant details for resolving their request (*InformIntent* and *InformSlot* respectively). Such acts provide a limited view of the actions taken by speakers in natural conversations, but do provide a way to identify and categorize automatable interactions in natural conversations. Table 1 provides an example conversation annotated with intents, slots, and dialogue acts.

4 Dataset Analysis

To better motivate NATCS as a proxy for natural, spoken form customer service conversations from multiple domains with a diverse set of intents, we compare with real conversations from commercial datasets comprising 5 call centers for retail and finance-related businesses (henceforth REAL). All datasets in REAL consist of manually-transcribed conversations between human agents and customers in live phone conversations where all personally-identifiable information has been pre-redacted. We restrict our analysis to datasets with primarily customer-initiated two-party dialogues.

As shown in Table 2, one surface-level distinction from publicly available TOD datasets is the average number of turns per conversation. Compared to MDGO, SGD, MWOZ and TM1$_{\text{Self}}$, REAL has considerably longer conversations (over 70 turns

	Collection	# Dialogues	# Turns/Conv	# Words/Turn	MTLD
TOD	MDGO	86,719	$15.9_{\pm 4.4}$	$11.9_{\pm 12.0}$	$46.6_{\pm 11.5}$
	MWOZ	10,437	$13.7_{\pm 5.2}$	$15.4_{\pm 7.4}$	$43.0_{\pm 9.5}$
	SGD	22,825	$20.3_{\pm 7.2}$	$11.7_{\pm 7.2}$	$38.3_{\pm 9.0}$
	TM1_{Self}	7,708	$22.0_{\pm 2.8}$	$10.3_{\pm 7.6}$	$45.9_{\pm 12.2}$
NATCS$_{\text{SELF}}$	Insurance	954	$70.6_{\pm 19.2}$	$12.3_{\pm 9.6}$	$45.4_{\pm 10.7}$
NATCS$_{\text{SPOKE}}$	Banking	980	$59.6_{\pm 23.1}$	$18.2_{\pm 19.4}$	$37.4_{\pm 6.7}$
	Finance	3,000	$65.6_{\pm 22.4}$	$16.3_{\pm 19.4}$	$36.9_{\pm 6.0}$
	Health	1,000	$67.0_{\pm 24.8}$	$16.3_{\pm 20.3}$	$34.7_{\pm 7.4}$
	Travel	1,000	$72.1_{\pm 24.7}$	$16.2_{\pm 20.0}$	$34.7_{\pm 7.1}$
REAL	Retail_{A}	4,500	$80.3_{\pm 41.7}$	$16.6_{\pm 14.9}$	$30.8_{\pm 4.8}$
	Retail_{B}	1,400	$52.8_{\pm 37.6}$	$17.9_{\pm 14.7}$	$32.8_{\pm 5.9}$
	Retail_{C}	4,500	$100.1_{\pm 69.9}$	$14.5_{\pm 12.7}$	$33.6_{\pm 5.5}$
	$\text{Finance}_{\text{A}}$	1,300	$61.7_{\pm 29.1}$	$17.6_{\pm 15.0}$	$38.4_{\pm 6.9}$
	$\text{Finance}_{\text{B}}$	1,700	$69.7_{\pm 43.8}$	$16.1_{\pm 13.5}$	$38.0_{\pm 5.9}$

Table 2: Comparison of dialogue datasets and corresponding high-level data characteristics. Task-oriented dialogue (TOD) datasets include MDGO (Peskov et al., 2019), MWOZ (Budzianowski et al., 2018), SGD (Rastogi et al., 2020), and TM1_{Self} (Byrne et al., 2019). We compare datasets collected from our two methodologies (NATCS$_{\text{SELF}}$ and NATCS$_{\text{SPOKE}}$), as well as 5 call center datasets (REAL). MTLD is a lexical diversity measure (McCarthy and Jarvis, 2010) computed at the conversation level.

per conversation on average, vs. 22 for TM1_{Self}). Furthermore, each turn has more words per turn, suggesting increasing complexity in spoken H2H dialogues. NATCS closely matches REAL in terms of conversation and turn lengths.

4.1 Intents and Slots

Table 3 provides a comparison of intent and slot annotations between existing synthetic datasets, NATCS and REAL. Datasets in REAL contain considerably more intents and slots for a particular domain than existing TOD datasets like MDGO. Turns containing intents are longer for both NATCS and REAL than SGD and MDGO.

Figure 2 compares the intent and slot distributions between Retail_{A} and NATCS$_{\text{SELF}}$ Insurance, indicating that both have skewed, long-tailed distributions of intents/slots, a product of the open intent/slot schemas used in NATCS.

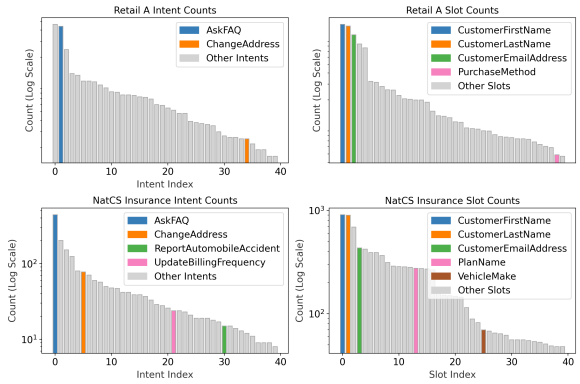

Figure 2: Intent and slot counts (logarithmic scale) for Retail_{A} and NATCS Insurance. We observe intents and slots in real data follow a Zipfian distribution in both REAL and NATCS.

4.2 Diversity Metrics

As we expect conversations in REAL to be less homogeneous than synthetic dialogue datasets, we compute automatic metrics to measure multiple aspects of diversity and compare with NATCS.

Conversational Diversity In Table 3, we examine the diversity of conversation flows as measured by the ratio of unique sequences of slots informed by the customer to the total number of sequences (e.g. slot *bigrams* or *trigrams*). In SGD, which constructs dialogue templates using a simulator, we observe a much lower percentage of unique n-grams than in REAL, despite SGD containing dialogues spanning multiple domains and services. On the other hand, while NATCS has lower slot n-gram diversity than REAL, both collection types have substantially higher slot n-gram diversity scores than both MDGO or SGD.

The average perplexity of a language model provides one indication of the difficulty of modeling

Collection	MDGO	SGD	NATCS$_{SELF}$	NATCS$_{SPOKE}$	REAL
Intents	6.8	46.0	61.0	94.0	64.0
Turns/Intent	473.1	1,235.5	42.5	27.3	47.4
Intent Turn Len.	8.2	14.1	17.8	29.8	23.9
Intents/Conv.	1.3	2.5	2.7	2.4	3.3
Sem. Diversity	21.3	28.6	34.5	31.1	43.5
Slots	13.8	112.0	246.0	238.0	150.0
Slots/Conv.	3.7	5.8	13.5	17.5	9.3
Slot 2-Gram %	2.8	1.9	16.3	12.4	24.2
Slot 3-Gram %	10.9	12.4	35.2	31.0	57.7

Table 3: Intent and slot annotation statistics for REAL, NATCS, MDGO, and SGD. *Sem. Diversity* is an sentence embedding-based metric for measuring intent-level diversity reported in Casanueva et al. (2022). *Slot n-gram %* indicates the ratio of unique sequences of slot annotations in conversations to the total number of such sequences.

Collection	PPL	PPL (ZS)
MDGO	4.9	10.7
MWOZ	6.8	14.6
SGD	7.0	10.1
TM1$_{Self}$	10.4	17.0
NATCS$_{SELF}$	11.6	15.7
NATCS$_{SPOKE}$	11.7	16.7
REAL	15.5	22.4

Table 4: Perplexity of GPT-Neo 125M on datasets in both fine-tuning (PPL) and zero-shot (ZS PPL) settings.

the dialogue in a given dataset (Byrne et al., 2019). High perplexity indicates higher difficulty, while low perplexity can indicate more uniform or predictable datasets. We compare between fine-tuning and zero-shot language modeling settings using GPT-NEO (Black et al., 2021). Zero-shot evaluation gives an indication of how compatible the dataset is with the pre-trained model without any fine-tuning. Section A.1 provides details on the fine-tuning procedure.

As shown in Table 4, we observe high perplexity on real data and low perplexity on synthetic datasets like SGD and MDGO. NATCS has lower perplexity than REAL, but considerably higher perplexity than MDGO, MWOZ, and SGD. Interestingly, there is a wider range of perplexities across real datasets, while most existing TOD datasets have a perplexity of 10 or less.

Intent Diversity We also investigate the semantic diversity of intent turns (SemDiv$_{intent}$) following Casanueva et al. (2022). Details of this calculation are provided in Section A.1. As shown in Table 3, we observe the highest SemDiv$_{intent}$ for REAL, but NATCS has considerably higher SemDiv$_{intent}$ than pre-existing synthetic datasets, indicating greater potential modeling challenges. We also compare the semantic of diversity of NATCS with other datasets for specific aligned intents, like *CheckBalance* in Appendix Table 9, also observing higher semantic diversity as compared to pre-existing intent classification benchmarks. Further investigation into the lexical diversity of intents is provided in Section A.1.

NATCS Dialogue Acts Distribution To better understand the characteristics of typical H2H customer service dialogues in call centers, we annotate a subset of real conversations with the "task-oriented" dialogue acts described in Section 3.2. Because NATCS dialogue acts consist of a small set of intent and slot-related functions commonly employed in automated TOD systems, we do not expect the labels to have high coverage in real conversations that do not revolve around a fixed set of intents and slots. However, they aim to provide a mechanism for aligning turns from natural conversations onto these automatable TOD constructs.

Table 5 compares the percentage of turns labeled with NATCS dialogue acts across multiple datasets along with the average counts of each label per dialogue. As expected, compared to synthetic data, dialogues in REAL have fewer turns labeled with NATCS dialogue acts (27.7%). More reflective of REAL in this regard, in NATCS$_{SPOKE}$ and NATCS$_{SELF}$, we observe that more than half of the turns in each conversation are not labeled, despite having higher total counts of dialogue acts per con-

Collection	Total (%)	InformIntent	ElicitSlot	ElicitIntent	InformSlot	ConfirmSlot
MDGO*	55.4	$1.1_{\pm 0.8}$	$2.7_{\pm 1.3}$	$2.0_{\pm 0.8}$	$3.1_{\pm 1.4}$	$0.7_{\pm 0.9}$
MWOZ*	70.5	$3.5_{\pm 1.9}$	$1.8_{\pm 1.4}$	$2.1_{\pm 1.4}$	$3.4_{\pm 2.1}$	$2.0_{\pm 1.4}$
$TM1_{Self}$*	52.9	$2.4_{\pm 1.7}$	$3.3_{\pm 1.8}$	$1.0_{\pm 1.1}$	$4.1_{\pm 1.7}$	$2.4_{\pm 1.7}$
SGD	60.6	$4.5_{\pm 2.3}$	$2.8_{\pm 1.7}$	$1.3_{\pm 1.1}$	$5.3_{\pm 2.3}$	$6.2_{\pm 4.1}$
$\text{NATCS}_{\text{SELF}}$	42.9	$4.5_{\pm 2.8}$	$9.2_{\pm 3.6}$	$3.1_{\pm 1.2}$	$11.8_{\pm 4.7}$	$8.2_{\pm 5.6}$
$\text{NATCS}_{\text{SPOKE}}$	44.7	$3.5_{\pm 2.1}$	$9.8_{\pm 5.2}$	$3.3_{\pm 1.5}$	$13.5_{\pm 6.8}$	$7.4_{\pm 5.9}$
REAL	27.7	$4.3_{\pm 3.1}$	$5.2_{\pm 3.9}$	$2.0_{\pm 1.1}$	$7.3_{\pm 5.4}$	$4.3_{\pm 4.9}$

Table 5: Percentage of turns containing NATCS dialogue acts in synthetic and real dialogue datasets. *Indicates distribution estimated from automatic predictions. A lower percentage of task-oriented turns is observed in NATCS and REAL than previous task-oriented dialogue datasets.

versation as compared to pre-existing datasets.

4.3 Human Evaluation

We also perform a human evaluation comparing MDGO, SGD, and REAL datasets against both $\text{NATCS}_{\text{SELF}}$ and $\text{NATCS}_{\text{SPOKE}}$. Rather than compare complete dialogues, because of the large disparity in conversation lengths, we restrict evaluation only to snippets including the first 5 turns after an intent is stated (including the turn containing the intent). Conversation snippets are graded on a scale of 1 to 5 along multiple dimensions, including **realism** (believability of the dialogue), **concision** (conciseness of customer, lack of verbosity), and **spoken-likeness** (possibility of being part of a spoken conversation). See Section A.2 Figure 3 for explicit definitions provided to graders.

The evaluation is conducted by 6 dialogue systems researchers, with each grader rating 50 randomly-selected conversations. As indicated by results in Table 6, conversations from REAL observe high values for both realism and spoken-likeness (4.71 and 4.95 respectively), with lower values for concision, indicating greater customer verbosity in real dialogues. The results also indicate that although expert human graders can still differentiate NATCS from real conversations, NATCS is graded as significantly more realistic and indicative of spoken modality than SGD and MDGO (two-tailed T-test with $p < 0.005$).

Dataset	Realism	Concision	Spokenlike
MDGO	2.85	4.21	3.18
SGD	3.36*	4.66**	3.18
$\text{NATCS}_{\text{SELF}}$	4.06**	3.88**	4.17**
$\text{NATCS}_{\text{SPOKE}}$	4.34	3.58	4.53
REAL	4.71*	2.87**	4.95**

Table 6: Human ratings comparing conversation snippets between REAL, NATCS and other task-oriented dialogue datasets along multiple dimensions (see Section A.2). Significance is based two-tailed t-tests ($^{*}p < 0.05$, $^{**}p < 0.005$), comparing samples in adjacent rows. For MDGO, we observe no significant difference for Concision ($p = 0.0522$) when comparing with $\text{NATCS}_{\text{SELF}}$, but do for $\text{NATCS}_{\text{SPOKE}}$ ($p = 0.0013$).

5 Applications

In this section, we investigate two potential applications of NATCS as a resource for building and evaluating systems related to human-to-human (H2H) customer service interactions. One goal of NATCS is to encourage research in the under-explored space of H2H task-oriented interactions, so these are intended to serve as motivating examples rather than prescribed uses.

5.1 NATCS as Training Data

One goal of this work is to accelerate the development of dialogue systems based on H2H conversations. While most existing work in intent induction assumes that customer turns corresponding to requests have already been identified, NATCS dialogue acts provide a mechanism to map turns onto TOD constructs like intents.

To validate the usefulness of dialogue act annotations in NATCS, we compare the cross-dataset generalization of dialogue act classifiers trained on annotations in NATCS against that of SGD, a large multi-domain corpus of task-oriented dialogues, evaluating on real conversations between human agents and customers.

We fine-tune ROBERTA-BASE using per-label binary cross entropy losses to support multiple la-

Train	P ↑	R ↑	F_1 ↑
SGD	$40.8_{\pm 1.6}$	$26.1_{\pm 3.1}$	$31.6_{\pm 2.7}$
NATCS	$64.6_{\pm 2.3}$	$46.8_{\pm 3.1}$	$54.1_{\pm 1.8}$
Real_{OOD}	$63.3_{\pm 3.2}$	$57.2_{\pm 2.9}$	$59.6_{\pm 0.6}$
Real_{ID}	$67.1_{\pm 2.4}$	$61.9_{\pm 2.5}$	$64.2_{\pm 0.5}$

Table 7: Comparison of dialogue act classifier performance on real datasets trained on SGD, NATCS, and in-domain (Real_{ID}) vs. out-of-domain (Real_{OOD}) real data. Training on NATCS achieves comparable precision to Real_{OOD}.

Turns	NMI	ACC	Purity	Inv. Purity
First	53.3	44.8	50.2	58.9
Pred.	56.7	49.1	55.7	59.9
Gold	61.6	49.4	63.9	55.0

Table 8: Intent clustering performance comparing different intent turn identification strategies (First, for first customer turns in each conversation, and Pred. for predicted dialogue acts) with using gold intent turns (Gold), averaged over Finance, Insurance, and Banking datasets in NATCS.

bels per sentence. Fine-tuning details are provided in Section A.3. We compare the dialogue act classification performance on real data when training on SGD, NATCS, and in-domain vs. out-of-domain real data. As shown in Table 7, a DA classifier trained on NATCS performs significantly better on real data than a classifier trained on SGD. Performance still lags behind that of training on real data, but with NATCS, the gap is closed considerably. In Section A.4, we also show that the dialogue act annotations in NATCS are able to generalize to new domains.

5.2 Intent Clustering with Noise

Recent work indicates growing interest in applications that can accelerate the development of these systems by automatically inducing TOD constructs such as intents and slots from customer support interactions (Yu et al., 2022; Shen et al., 2021; Kumar et al., 2022; Perkins and Yang, 2019; Chatterjee and Sengupta, 2020). To further motivate NATCS as a realistic test bed for applications that learn from natural conversations, we demonstrate how it can serve as a benchmark for unsupervised intent clustering tasks.

In a realistic setting, turns in conversations containing intents will not be provided in advance. We thus compare three settings: 1) using the first customer turn in each conversation 2) using turns predicted as having intents with a dialogue act classifier and 3) using turns labeled with intents (*gold* dialogue acts).

Utterances are encoded using a sentence embedding model from the SENTENCETRANSFORMERS library (Reimers and Gurevych, 2019), ALL-MPNET-BASE-V2. We use k-means clustering with the number of clusters set to the number of reference intents. To assign cluster labels to all gold input turns, we use label propagation by training a logistic regression classifier using ALL-MPNET-BASE-V2 embeddings as static features on inputs assigned cluster labels (such as first turns), then apply the classifier to the missing turns to get predicted cluster labels.

The results, shown in Table 8, demonstrate that using automatically predicted turns leads to a drop in purity and NMI. The drop in purity is attributable to irrelevant, non-intentful turns being clustered together with relevant intents, a potentially costly error in real-world settings that is not typically reflected in intent clustering evaluation.

6 Conclusions

We present NATCS, a corpus of realistic spoken human-to-human customer service conversations. The collection of NATCS is complexity-driven and domain-restricted, resulting in a dataset that better approximates real conversations than pre-existing task-oriented dialogue datasets along a variety of both automated and human-rated metrics. We demonstrate two potential downstream applications of NATCS, showing that training using NATCS results in better performance with real test data compared to training using other publicly-available goal-oriented datasets, and that NATCS can provide a new challenging benchmark for realistic evaluation of intent induction.

We hope that NATCS will help facilitate open research in applications based on customer support conversations previously accessible mainly in industry settings by providing a more realistic annotated dataset. In future work, we hope to expand on annotations in NATCS to support more tasks such as call summarization, response selection, and generation.

Limitations

NATCS is partially annotated with dialogue acts, intents, and slots, which are annotated independently from the initial collection of the conversations. While decoupling annotations from collection was intended to facilitate natural and diverse dialogues, the methodology is more time-consuming and expensive than previous approaches that use pre-structured conversation templates to avoid the need for manual annotation. In particular, $\text{NATCS}_{\text{SPOKE}}$ requires multiple participants engaging in synchronous conversations, followed by independent manual transcriptions and annotations, making the approach particularly time-consuming and difficult to apply for large collections. Furthermore, this decoupling of annotations from collection has greater potential for annotator disagreement.

While the complexity types and annotations are mostly language-agnostic, NATCS is restricted to EN-US customer-initiated customer service conversations between a single agent and customer in a limited number of domains (multi-party conversations beyond two participants or agent-initiated conversations are not included). The annotations included are primarily intended for applications related to task-oriented dialogue systems.

Further, we note that NATCS closes the gap from real conversations along many metrics, but still falls short along some dimensions. We find that real conversations are more verbose, more believable, and less predictable. We also note that comparisons in our paper focused on a limited number of task-oriented dialogue datasets with different collection approaches, and did not exhaustively include all pre-existing dialogue datasets for comparison.

Ethics Statement

In this paper, we present a new partially-annotated dataset. In adherence with the ACL code of conduct and recommendations laid out in Bender and Friedman (2018) it is appropriate to include a data statement. Our dataset is completely novel, and was collected specifically to support the development of natural language systems. Workers who are proficient in the EN-US variant of English were hired through a vendor with a competitive hourly rate compared to the industry standard for language consultants. For $\text{NATCS}_{\text{SPOKE}}$, these workers spoke to each other and then transcribed the data. For $\text{NATCS}_{\text{SELF}}$ these workers wrote the conversations.

To annotate the data, we used two pools of annotators. Both had formal training in linguistics and were proficient in the EN-US variant of English. One pool was hired through a vendor with a competitive hourly rate. The other pool consisted of full-time employees.

Curation Rationale Our dataset includes all of the data that was produced by the consultants we hired. Quality Assurance was done on a subset of this data. We hope that any concerns would have shown up in this sample. We annotated a random subset of the full dataset.

Language Variety The dataset is EN-US. The speakers (or writers) were all fluent speakers of EN-US. We did not target a particular sub-type of the EN-US language variety.

Speaker Demographics We do not have detailed speaker demographics, however, we do have male and female speakers from a variety of age ranges.

Annotator Demographics We do not have detailed annotator demographics, however, we do have male and female speakers from a variety of age ranges. All annotators had at least some formal linguistics training (ranging from a B.A. to a Ph.D.).

Speech Situation For $\text{NATCS}_{\text{SPOKE}}$, speakers were talking in real time on the phone to one another. It was semi-scripted. Speakers were not told exactly what to say, but were given some constraints.

References

Amazon Contact Lens. 2023. Contact Lens. https://aws.amazon.com/connect/contact-lens/.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 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, Brussels, Belgium. Association for Computational Linguistics.

Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Ben Goodrich, Daniel Duckworth, Semih Yavuz, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik. 2019. Taskmaster-1: Toward a realistic and diverse dialog dataset. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4516–4525, Hong Kong, China. Association for Computational Linguistics.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 38–45, Online. Association for Computational Linguistics.

Inigo Casanueva, Ivan Vulić, Georgios Spithourakis, and Paweł Budzianowski. 2022. NLU++: A multi-label, slot-rich, generalisable dataset for natural language understanding in task-oriented dialogue. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1998–2013, Seattle, United States. Association for Computational Linguistics.

Ajay Chatterjee and Shubhashis Sengupta. 2020. Intent mining from past conversations for conversational agent. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4140–4152, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Derek Chen, Howard Chen, Yi Yang, Alexander Lin, and Zhou Yu. 2021. 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*, pages 3002–3017, Online. Association for Computational Linguistics.

Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. 2017. Frames: a corpus for adding memory to goal-oriented dialogue systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 207–219, Saarbrücken, Germany. Association for Computational Linguistics.

Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D. Manning. 2017. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 37–49, Saarbrücken, Germany. Association for Computational Linguistics.

Google Contact Center AI. 2023. Contact Center. https://cloud.google.com/solutions/contact-center.

Timothy Hewitt and Ian Beaver. 2020. A case study of user communication styles with customer service agents versus intelligent virtual agents. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 79–85, 1st virtual meeting. Association for Computational Linguistics.

John F Kelley. 1984. An iterative design methodology for user-friendly natural language office information applications. *ACM Transactions on Information Systems (TOIS)*, 2(1):26–41.

Rajat Kumar, Mayur Patidar, Vaibhav Varshney, Lovekesh Vig, and Gautam Shroff. 2022. Intent detection and discovery from user logs via deep semi-supervised contrastive clustering. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1836–1853, Seattle, United States. Association for Computational Linguistics.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. 2019. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1311–1316, Hong Kong, China. Association for Computational Linguistics.

Q Vera Liao, Biplav Srivastava, and Pavan Kapanipathi. 2017. A measure for dialog complexity and its application in streamlining service operations. *arXiv preprint arXiv:1708.04134*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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

Microsoft Digital Contact Center Platform. 2023. Microsoft Digital Contact Center Platform. https://www.microsoft.com/en-us/microsoft-cloud/contact-center.

Hugh Perkins and Yi Yang. 2019. Dialog intent induction with deep multi-view clustering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*

(EMNLP-IJCNLP), pages 4016–4025, Hong Kong, China. Association for Computational Linguistics.

Denis Peskov, Nancy Clarke, Jason Krone, Brigi Fodor, Yi Zhang, Adel Youssef, and Mona Diab. 2019. Multi-domain goal-oriented dialogues (MultiDoGO): Strategies toward curating and annotating large scale dialogue data. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4526–4536, Hong Kong, China. Association for Computational Linguistics.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8689–8696.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Xiang Shen, Yinge Sun, Yao Zhang, and Mani Najmabadi. 2021. Semi-supervised intent discovery with contrastive learning. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 120–129, Online. Association for Computational Linguistics.

Tsung-Hsien Wen, David Vandyke, Nikola Mrksic, Milica Gasic, Lina M Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2016. A network-based end-to-end trainable task-oriented dialogue system. *arXiv preprint arXiv:1604.04562*.

Dian Yu, Mingqiu Wang, Yuan Cao, Izhak Shafran, Laurent Shafey, and Hagen Soltau. 2022. Unsupervised slot schema induction for task-oriented dialog. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1174–1193, Seattle, United States. Association for Computational Linguistics.

A Appendix

A.1 Experimental Setting Details

In this section, we provide details on the experimental settings used for evaluation of NATCS and other dialogue datasets.

Perplexity Evaluation To evaluate the language modeling (LM) perplexity on each dataset, we compare both a fine-tuning setting as well as a zero-shot setting using GPT-NEO (Black et al., 2021) as the pre-trained LM. We fine-tune GPT-NEO on each dataset, sampling 4096 blocks of 128 tokens as training data and evaluating on held-out test splits. Fine-tuning is performed for 6 epochs with a batch size of 64 and learning rate of 5e-5. Perplexity is computed at the level of bytes using a sliding window of 128 tokens.

Semantic Diversity Evaluation Following Casanueva et al. (2022), to compute semantic diversity for a single intent, we (1) compute intent centroids as the average of embeddings for the turns labeled with the intent using the Sentence-BERT (Reimers and Gurevych, 2019) library with the pre-trained model ALL-MPNET-BASE-V2, then (2) find the average cosine distance between each individual turn and the resulting centroid. Finally, (3) overall semantic diversity scores ($\text{SemDiv}_{\text{intent}}$) in Table 3 are computed as a frequency-weighted average over intent-level scores.

Table 9 shows the semantic diversity scores (Casanueva et al., 2022) for different intents aligned across several datasets.

Lexical Diversity of Intents As an additional indicator of intent-level diversity, we measure the frequency-weighted average of type token ratios for utterances within each intent ($\text{TTR}_{\text{intent}}$). To account for the redaction of names and numbers in real data, we perform a similar redaction step on all datasets, automatically converting names and numbers to a single PII token with regular expressions and a named entity tagger before computing TTR. Shown in Table 10, we observe similar $\text{TTR}_{\text{intent}}$ between NATCS and REAL, while the pre-existing synthetic datasets lag behind considerably.

A.2 Human Evaluation Details

Guidelines provided to human graders are provided in Figure 3. **Realism** measures the overall believability that the conversation could have taken place in a real scenario (which penalizes unlikely, silly utterances from the customer or unprofessional behavior from the agent). **Concision** measures how concise the customer responses are, with lower scores for lengthy utterances containing details that are unnecessary for resolving their request. **Spoken-like** measures the likelihood that the conversation was originally spoken in a phone conversation (as opposed to being written from a chatroom or messaging platform).

50 of the 300 conversation snippets evaluated were graded by pairs of annotators (5 pairs total).

Intent	NATCS	MultiDoGO	CLINC150	BANK77	SGD
CheckBalance	**31.9**	17.9	27.8	_	23.1
MakeTransfer	**34.3**	24.2	29.5	_	25.9
ReportLostStolenCard	**29.0**	18.6	16.2	18.4	_
DisputeCharge	**35.3**	23.7	26.1	_	_
OrderChecks	**31.8**	21.5	19.0	_	_
CloseBankAccount	**26.4**	17.6	_	20.1	_
UpdateStreetAddress	**31.4**	17.5	_	28.6	_
ChangePin	**27.4**	_	20.3	19.7	_

Table 9: Comparing semantic diversity (Casanueva et al., 2022) for various aligned intents across MultiDoGO (Peskov et al., 2019), CLINC150 (Larson et al., 2019), BANK77 (Casanueva et al., 2020), and SGD (Rastogi et al., 2020).

Collection	MDGO	SGD	$\text{NATCS}_{\text{SELF}}$	$\text{NATCS}_{\text{SPOKE}}$	REAL
1-Gram $\text{TTR}_{\text{intent}}$	7.2	2.5	16.7	17.0	16.0
2-Gram $\text{TTR}_{\text{intent}}$	21.0	13.7	48.1	53.1	56.5
3-Gram $\text{TTR}_{\text{intent}}$	33.0	28.6	66.7	76.6	83.2

Table 10: $\text{TTR}_{\text{intent}}$ provides intent-level type-token-ratios after removing names and numbers. Type-token-ratios are computed as 1-grams, 2-grams, and 3-grams, for which we observe consistently higher diversity for NATCS and REAL than MDGO and SGD.

For these, we observed a Krippendorff's alpha of 0.52, 0.59, and 0.37 respectively for Spoken-like, Realism, and Concision respectively.

A.3 Dialogue Act Classifier Training Details

The dialogue act classifier is implemented by fine-tuning ROBERTA-BASE (Liu et al., 2019) using per-label binary cross entropy losses to support multiple labels per sentence. To encode dialogue context, we append the three previous sentences to the current turn with a separator token ($[SEP]$), adding a speaker label to each sentence and using padding tokens at the beginning of the conversation. Fine-tuning is performed for 6 epochs with a batch size of 16, using AdamW with a learning rate of 2e-5. Dataset-level dialogue act classifier performance is provided in Table 11.

A.4 Dialogue Act Cross-Domain Generalization

For models trained on NATCS to be useful in practice, they must be able to generalize beyond the limited number of domains present in NATCS. To measure cross-domain generalization, we perform cross validation, training separate models in which we hold out a single domain (e.g. train on Banking, evaluate on Travel). As shown in Table 12, performance on the heldout domains is lower, particularly for recall, but does not drop substantially.

A.5 Collection Methodology

Table 13 provides an example company profile used to assist in achieving cross-dataset consistency. Table 14 provides an example intent/slot schema for a ResetPassword intent in the Insurance domain. Table 15 provides example utterances for each discourse complexity. Table 16 provides examples of spoken complexities used to simulate spoken-form conversations in the Insurance.

Figure 4 provides target percentages for complexities used to simulate spoken conversations in $\text{NATCS}_{\text{SELF}}$.

A.6 Conversation Examples

Tables 17, 18, and 19 provide examples of conversations collected in $\text{NATCS}_{\text{SPOKE}}$. Tables 20, 21, and 22 provide examples of conversations collected in $\text{NATCS}_{\text{SELF}}$.

Read each conversation excerpt, then fill in values for each of the below dimensions with ratings between 1 and 5.

Realism – Is the customer making a realistic request? Is it believable that the conversation took place in a real setting? Is the customer's request or problem something that you imagine customers can encounter, or is it something incredibly unlikely or silly? Does the agent respond in a professional manner? Is it believable that the conversation took place in real life?

- 5 (realistic) means that the conversation is realistic and could have taken place in a real life setting.
- 1 (unrealistic) means that it is highly unlikely that the conversation could have taken place in a real life setting.

Concision – Does the customer provide only the necessary details for resolving their request? Does the customer avoid providing long responses or extra details that aren't 100% relevant to the conversation or for resolving their request?

- 5 (concise) means that the customer was consistently concise and clear with their responses to the agent.
- 1 (not concise) means that the customer provides multiple extra unnecessary details to the agent and provides long responses.

Spoken-like – Does the excerpt appear to be from a spoken (as opposed to written) conversation? Are there signals that indicate that the conversation was originally spoken and then transcribed (as opposed to occurring over a chatroom or messaging platform)?

- 5 (spoken) means that it is extremely likely that the conversation was originally spoken.
- 1 (written) means that it is unlikely that the conversation was originally spoken (and was probably written as a chat conversation instead).

Figure 3: Guidelines used in human evaluation.

Test	**Train**	**P** ↑	**R** ↑	**F**$_1$ ↑
Finance$_{\text{A}}$	SGD	$41.0_{\pm 0.7}$	$30.9_{\pm 2.3}$	$35.2_{\pm 1.7}$
	NATCS	$67.1_{\pm 1.5}$	$45.7_{\pm 3.3}$	$54.3_{\pm 2.8}$
Finance$_{\text{B}}$	SGD	$42.5_{\pm 1.1}$	$29.8_{\pm 2.9}$	$35.0_{\pm 2.1}$
	NATCS	$59.5_{\pm 1.2}$	$43.5_{\pm 2.5}$	$50.3_{\pm 2.1}$
Retail$_{\text{A}}$	SGD	$42.5_{\pm 2.6}$	$20.3_{\pm 3.6}$	$27.4_{\pm 3.7}$
	NATCS	$65.4_{\pm 4.4}$	$47.3_{\pm 3.8}$	$54.7_{\pm 1.2}$
Retail$_{\text{B}}$	SGD	$37.2_{\pm 2.1}$	$23.3_{\pm 3.5}$	$28.6_{\pm 3.3}$
	NATCS	$66.1_{\pm 2.3}$	$50.5_{\pm 2.9}$	$57.1_{\pm 1.1}$
Average	SGD	$40.8_{\pm 1.6}$	$26.1_{\pm 3.1}$	$31.6_{\pm 2.7}$
	NATCS	$64.6_{\pm 2.3}$	$46.8_{\pm 3.1}$	$54.1_{\pm 1.8}$
	Real$_{\text{OOD}}$	$63.3_{\pm 3.2}$	$57.2_{\pm 2.9}$	$59.6_{\pm 0.6}$
	Real$_{\text{ID}}$	$67.1_{\pm 2.4}$	$61.9_{\pm 2.5}$	$64.2_{\pm 0.5}$

Table 11: Comparison of dialogue act classifier performance on real datasets trained on SGD, NATCS, and in-domain (Real$_{\text{ID}}$) vs. out-of-domain (Real$_{\text{OOD}}$) real data. Training on NATCS achieves comparable precision to Real$_{\text{OOD}}$.

Test	Train	P ↑	R ↑	F_1 ↑
Banking	ID	$84.7_{\pm 0.7}$	$85.3_{\pm 0.0}$	$85.0_{\pm 0.3}$
	OOD	$85.4_{\pm 1.5}$	$86.0_{\pm 1.1}$	$85.7_{\pm 1.2}$
Finance	ID	$84.2_{\pm 0.3}$	$85.3_{\pm 0.4}$	$84.8_{\pm 0.4}$
	OOD	$87.2_{\pm 0.8}$	$88.5_{\pm 0.5}$	$87.8_{\pm 0.2}$
Health	ID	$86.1_{\pm 0.8}$	$87.4_{\pm 0.6}$	$86.8_{\pm 0.1}$
	OOD	$80.4_{\pm 1.1}$	$78.3_{\pm 0.9}$	$79.3_{\pm 0.3}$
Insurance	ID	$84.8_{\pm 0.9}$	$86.3_{\pm 0.6}$	$85.6_{\pm 0.2}$
	OOD	$87.6_{\pm 0.2}$	$79.4_{\pm 0.6}$	$83.3_{\pm 0.3}$
Travel	ID	$85.9_{\pm 0.7}$	$85.1_{\pm 0.4}$	$85.5_{\pm 0.4}$
	OOD	$84.9_{\pm 1.0}$	$80.0_{\pm 0.8}$	$82.4_{\pm 0.3}$
Average	ID	$85.2_{\pm 0.7}$	$85.9_{\pm 0.4}$	$85.5_{\pm 0.3}$
	OOD	$85.1_{\pm 0.9}$	$82.4_{\pm 0.8}$	$83.7_{\pm 0.5}$

Table 12: Dialogue act classifier cross-domain evaluation on NATCS. In-domain (ID) training data consists of data from the same domain as the test dataset, whereas out-of-domain (OOD) training data consists of training data from the remaining (non-test) domains.

Spelling (80%), HesitationsFillersLow (60%), Backchanneling (50%), HesitationsFillersHigh (40%), Interruption (40%), Confirmation (30%), Disfluencies (30%), Rambling (30%), AskToRepeatOrClarify (25%), PartialInformation (10%)

Figure 4: Complexities present in spoken form conversations and corresponding target percentages, estimated based on manual inspection of a small set of real conversations.

Table 13: Sample company profile for NATCS(Insurance)

Company name	Rivertown Insurance
Domain	Insurance
Description	An ordinary insurance company. Provides insurance for Pets, Rent, Automobile, Life, etc.
Website	www.rivertowninsurance.com
Insurance offered	Life Pet Automobile Condo Homeowner Renter
Plan Types offered (Automobile)	Basic Auto ($1000/year) Preferred Auto ($1500/year) Complete Auto ($2000/year)
Plan Types offered (Condo)	Basic ($500/year) Condo Preferred ($600/year)
Plan Types offered (Homeowner)	Basic Home ($1200/year) Home Preferred ($1600/year) Home Complete ($2000/year)
Plan Types offered (Life)	Term Life Insurance ($300/year) Whole Life Insurance ($1800/year) Universal Life Insurance ($1200/year)
Plan Types offered (Pet)	Petcare Basic ($500/year) Petcare Preferred ($1000/year)
Plan Types offered (Renters)	Renters Basic ($200/year) Renters Preferred ($300/year)
Security questions:	What is your mother's maiden name? What is the name of your childhood best friend? What is the name of your high school? What is the name of your first pet? What is the name of your favorite teacher?
Company's protocol to verify identity	To verify a customer's identity, you will need either: 1) FirstName, LastName, DateOfBirth, and CustomerNumber (8 digits long), or 2) FirstName, LastName, DateOfBirth, PhoneNumber, SocialSecurityNumber, and answer to one security question

Slots	Agent suggested talking points	Customer suggested talking points
	1. Confirm the identity of the customer 2. Ask if there's anything else you can do for the customer before completing the conversation.	1. Ask if you can change the password on the website instead of making a phone call
Slots required for identity verification	To verify a customer's identity, you will need either: 1) FirstName, LastName, DateOfBirth, and CustomerNumber (8 digits long), or 2) FirstName, LastName, DateOfBirth, PhoneNumber, SocialSecurityNumber, and answer to security question	
SecurityQuestionAnswer		
EmailAddress	1. Verify the email address on file. Note that the reset link will be sent to this email	

Table 14: Sample intent/slot schema with instructions for Insurance, ResetPlan intent

Table 15: Discourse complexities with examples

Name	Description	Example
ChitChat	Small talk unrelated to the intent	Hey, how's your day going?
Frustration	Expression of frustration	That's such a scam because you were able to fulfill this before
BackgroundDetail	Related information that is unnecessary for resolving the request is provided for context	**We booked a room from you last year** and were hoping to come stay again over Memorial day weekend
ImplicitDescriptiveIntent	A description of the problem is provided, rather than an explicit request	I got this notice in the mail regarding a rate increase but an agent told me that the rate was fixed till the end of the year
SlotChange	The customer amends or corrects information provided	Oh wait, my zipcode is actually 20512 now
IntentChange	The customer changes their request midway through a conversation.	Actually, I'd better check my balance first
FollowUpQuestion	The customer asks follow-up questions that are related to their original intent (can occur before the original intent has been fulfilled)	How long does it take to deliver?
MultiElicit	The agent elicits multiple related pieces of information at once	Could you please provide your name and date of birth?
MultiIntent	The customer has multiple requests	I'd also like to transfer some money while I'm here
MultiIntentUpfront	The customer states multiple requests at the beginning of the conversation	I wanted to check the status of a claim and add my wife to my policy
MultiValue	The customer provides multiple slot values, even if only a single value was requested	A: What item are you returning today? / C: I wanted to return a blender and a pan
Callback	The conversation is a continuation of another conversation with prior context	Hey I'm back now. I was able to restart
MissingInfoWorkaround	The agent requests other information because the customer does not have some requested information	C: I don't have the barcode number / A: No worries, could you give me your last name?
MissingInfoCantFulfill	The agent is uable to fulfill a request because the customer is missing some information	C: I don't have the barcode number / A: I will need that number to fulfill this request for you, so you'll need to call back once you find it
PauseForLookup	Customer or agent asks the other party to hold or wait while they look up some information.	Can you hold on while I find that number?
SlotLookup	The agent already has information about the customer and only needs to verify details	Is johndoe@gmail.com still a good email address for you?
Overfill	The customer provides more information than asked	A: What's your first name? / C: It's John Doe, email is johndoe@gmail.com
SlotCorrection	The agent corrects a customer regarding a product or service offered by the company	C: I ordered the Derek pot set / A: The Darin set, yes I see your order here.

Table 16: Spoken complexities with examples

Name	Description	Example
Backchanneling	Speaker(s) interrupts to signify that they are listening	uh huh, right go on
Confirmation	Speaker(s) confirms their understanding of what the other party says, either due to being unable to hear properly or as an acknowledgement	you said thirty dollars for everything?
Disfluencies	Speaker(s) exhibit disfluencies and unfinished thoughts	No no that it isn't it. You'll need to get the um first can you see in the right hand corner your username?
HesitationsFillersHigh	Speaker(s) includes a high percentage of filler words to mark a pause or hesitation	um, uh, er
HesitationsFillersLow	Speaker(s) includes a low percentage of filler words to mark a pause or hesitation	um, uh, er
Interruption	Speaker(s) interrupts, especially when it is unclear when a thought has ended	C: Yeah I wanted to get help with filling out these forms and then um / A: Okay sure I can help with that / C: I'm also going to want to change some personal information
PartialInformation	Speaker(s) is only able to give partial information due to forgetfulness	I'm trying to get a hold of that thing, you know the 228 something form
Rambling	Speaker(s) ramble and repeat themselves. They may paraphrase themselves	Oh I got it I see. So basically I had this employee who just up and moved and it's been a disaster so I don't know um like where to even find him, but he was in charge of all the records so things are just a huge mess and I don't know how to update this it's been impossible and I don't really understand everything that goes into this.
AskToRepeatOrClarify	One party is unable to hear and either expresses this or asks for repetition or clarification.	Did you say that's 128 or 1228?
Spelling	Speakers should spell things out if they need to clarify uncommon or ambiguous spellings	That's Jon spelled without an H.

Text	Annotations	Complexities
A: Good afternoon. Welcome to Intellibank. This is Rose. How can I help you today?	ElicitIntent	
C: Hi Rose, my name is Kim$_{\text{FIRSTNAME}}$. I have a couple of questions for you actually I need to make a wire transfer and I also have an issue with my card. I may have lost it. So, but I'd like to to do the wire transfer first and then well maybe we can re-issue me a new card.	InformSlot, InformIntent (ExternalWireTransfer, ReportLostStolenCard, RequestNewCard)	MultiIntentUpfront, Overfill
A: OK.		Backchanneling
A: OK, certainly. I can help you with that issue. Kim, first I need to know your last name$_{\text{LASTNAME}}$.	ElicitSlot	
C: Sure, it's Johns$_{\text{LASTNAME}}$, J O H N S$_{\text{LASTNAME}}$.	InformSlot	Spelling
A: All right, thank you so much Ms. Johns. And also, what is your date of birth$_{\text{DATEOFBIRTH}}$, please?	ElicitSlot	
C: It's April twenty-fourth nineteen seven seven$_{\text{DATEOFBIRTH}}$.	InformSlot	
A: OK, thank you so much. and I also need the account number$_{\text{ACCOUNTNUMBER}}$ that you would like to make the transfer from.	ElicitSlot	
C: sure, hold on for one moment. Let me go grap that.		PauseForLookup
A: Sure, take your time.		
C: All right, I have it right here now.		
...		

Table 17: Example conversation from NATCS Banking. The initial customer utterance is labeled with multiple intents (ExternalWireTransfer, ReportLostStolenCard, and RequestNewCard).

Text	Annotations	Complexities
A: Thank you for calling Intellibank. My name is Izumi. Who$_{\text{UNSPECIFIED}}$ am I speaking with today?	ElicitSlot	
C: Hey, Izumi. My names Edward$_{\text{FIRSTNAME}}$ Elric$_{\text{LASTNAME}}$. just calling in cuz I have a big problem today. I was at the I was at a couple of stores with my wife buying supplies for a house and I think I lost my card. I can't remember when I lost it since we were at one of the stores. My wife actually used one of her cards to get the store points that they offer. don't know if it was before then or after. I think it was after because we went to go eat somewhere there. just calling in to get some help.	InformIntent, InformSlot (ReportLostStolenCard)	BackgroundDetail, ImplicitDescriptiveIntent
A: Oh, no. I'm so sorry to hear that Mr. Elric. I know that it's difficult when you lose your your bank card. but yeah. Please rest assured, I will do everything I can to make sure that your account is secure and then we'll get you a new card as well as soon as possible, OK?		ChitChat
C: Oh, sweet. Never lost my card before. Kind of worried. So not sure what I have to do. Is there anything that I need to give you first so I can get my card?	InformIntent	Frustration, FollowUpQuestion
A: Yes, Mr. Elric. So first of all, I'll just need to ask for your personal information so that I can pull up your account . can I get your account number$_{\text{ACCOUNTNUMBER}}$, please?	ElicitSlot	
C: Yeah. It is one zero zero, one seven zero$_{\text{ACCOUNTNUMBER}}$ and the last four of my social are fifty-two fourteen$_{\text{LASTFOURSOCIAL}}$ in case you need that.	InformSlot	Overfill
A: yes. Thank you for that information. And then, I just wanna make sure that I have that account number correct. So I have one zero zero, one seven zero$_{\text{ACCOUNTNUMBER}}$. And then, the last four of the social were five two one four$_{\text{LASTFOURSOCIAL}}$.	ConfirmSlot	
C: Yeah yeah. That's right.		
A: Perfect. And then, can I also verify your date of birth$_{\text{DATEOFBIRTH}}$, please?	ElicitSlot	
C: Yeah. It is October fourth, nineteen ninety$_{\text{DATEOFBIRTH}}$.	InformSlot	
A: OK, perfect. Thank you so much. Let me just pull up your account. Give me one moment.		PauseForLookup
C: Oh, OK.		
A: All right. Just one second.		PauseForLookup
C: Yeah. I really hope get help to find my card. We are renovating our house at the moment right now. Started redoing our walls not too long ago. It's a bunch of wallpaper, so we just need help finishing removing it. And then, my wife is gonna head off to the store to get some paint to start that project.		BackgroundDetail, ChitChat
A: Oh, that's awesome. Do you guys have like a color scheme or color palette that you're working with?		ChitChat
...		

Table 18: Example conversation from NATCS Banking with discourse complexities (e.g. BackgroundDetail, ChitChat, and FollowUpQuestion). In this conversation, the customer provides considerable background information related, but not necessary for understanding their intent.

Text	Annotations	Complexities
A: Hello and thank you for calling Intellibank. This is Mark speaking. How could I help you today?	ElicitIntent	
C: Hi Mark. My name is Dorothy$_{\text{FirstName}}$ Lee$_{\text{LastName}}$. I would like to check my savings$_{\text{TypeOfAccount}}$ account balance.	InformSlot, InformIntent (CheckAccountBalance)	Overfill
A: OK Dorothy. I can help you with that if you give me one second. Could you in the meantime give me your date of birth$_{\text{DateOfBirth}}$ please?	ElicitSlot	
C: Yeah. It's April the fifteenth nineteen ninety-nine$_{\text{DateOfBirth}}$.	InformSlot	
A: OK perfect. Thank you so much. Now also could you give me your account number$_{\text{AccountNumber}}$?	ElicitSlot	
C: Oh no I don't have it with me. Is that a problem?		
A: That shouldn't be a problem. do you happen to have your credit card number$_{\text{CreditCardNumber}}$ on you as well?	ElicitSlot	MissingInfoWorkaround
C: I don't have that either.		
A: OK well I need some form of identifying you. do you happen to have a driver's license$_{\text{OtherIDNumber}}$ or another state ID issued number?	ConfirmSlot, ElicitSlot	MissingInfoWorkaround
C: No sorry I I didn't bring anything with me today. Is there any other way we can do it?	InformIntent	Disfluencies
A: unfortunately I'm gonna need some of that information to to process your request. so unfortunately because there's a lot of of theft going on I I it's could be fraud. I'm not sure that you are who you say who you are and if you can't give me that information. We use those as security checkpoints then I won't be able to complete your request for you. I apologize for that Mrs. Dorothy Lee.		MissingInfoCantFulfill, Disfluencies
C: Well I'm kind of disappointed because I always get like a terrible service customer here but OK. I want you to help me with another thing. Is that possible?		Frustration, MultiIntent
A: Yes of course. what what else could I help you with today? .	ElicitIntent	
...		

Table 19: Example conversation from NATCS Banking with discourse complexities (e.g. MultiIntent and MissingInfoWorkAround). In this conversation, the customer is unable to provide the necessary information for identity verification, despite the agent offering multiple possible workarounds.

Text	Annotations	Complexities
A: Thank you for calling Rivertown Insurance. How may I help you today?	ElicitIntent	
C: Yes. I need to do something about lowering my premiumiss I recently lost my job and just can't afford the payments anymore. I don't want to change companies. I've been with you guys for years.	InformIntent (ChangePlan, RequestDiscount)	BackgroundDetail
A: I'm sorry to hear you lost your job. Let's see what we can do.		
C: Okay. Thank you.		
A: May I have your first$_{\text{FirstName}}$ and last name$_{\text{LastName}}$?	ElicitSlot	MultiElicit
C: It's Maria$_{\text{FirstName}}$ Sanchez$_{\text{LastName}}$.	InformSlot	
A: Thank you, Maria. Do you happen to have your customer number$_{\text{CustomerID}}$?	ConfirmSlot, ElicitSlot	
C: I think so. Let me check my purse.		PauseForLookup
A: Okay. Take your time.		
C: It's one two three four five six seven eight$_{\text{CustomerID}}$.	InformSlot	
A: Perfect, and can you verify your date of birth$_{\text{DateOfBirth}}$ please?	ElicitSlot	
C: It's seven twenty six nineteen eighty nine$_{\text{DateOfBirth}}$.	InformSlot	
A: Thank you, Maria. I have you pulled up here. Which$_{\text{PlanType}}$ policy were you looking at reducing the payment on? Life or Auto?	ConfirmSlot, ElicitSlot	SlotLookup
C: the auto policy$_{\text{PlanType}}$. I don't want to change my life insurance.	InformSlot	
A: Okay. It looks like you're on the Complete plan$_{\text{PlanName}}$. Does that sound correct?	ConfirmSlot, ElicitSlot	SlotLookup
C: Yes it was the highest one.		
A: Okay. We do have two options with lower payments how much lower did you need to go?	ElicitSlot	
...		

Table 20: Example conversation from NATCS$_{\text{SELF}}$ demonstrating primarily discourse complexities.

Text	Annotations	Complexities
A: Hello, Rivertown Insurance, Marissa here how could I help?	ElicitIntent	
C: Yep I just made my account and now it says call to enroll?	InformIntent (EnrollInPlan)	Callback, BackgroundDetail
A: Mhm I could enroll you in one of our plans over the phone.		
C: Uh-huh.		Backchanneling
A: Do you have any idea what sort of plan$_{\text{PLANTYPE}}$ you wanted to enroll i-in?	ElicitSlot	Disfluencies
C: Nah could I also get a quote over the phone?	InformIntent (GetAutoQuote)	MultiIntent
A: Mhm I could also do that for you sir.		
C: Oh great Marissa thank you.		
A: Mhm no problem sir, could I get your first and last name$_{\text{FIRSTNAME}}$first and last name$_{\text{LASTNAME}}$ p-please?	ElicitSlot	Disfluencies
C: Jakob$_{\text{FIRSTNAME}}$ Burbert$_{\text{LASTNAME}}$.	InformSlot	
A: Jacob j.a.c.o.b$_{\text{FIRSTNAME}}$?	ConfirmSlot	Spelling, Confirmation
C: Nah Jakob with a k$_{\text{FIRSTNAME}}$.	InformSlot	Spelling
A: Got it is Burbert b.u.r.b.e.r.t$_{\text{LASTNAME}}$?	ConfirmSlot	Spelling, Confirmation
C: Yep!		
A: Oh great so I'm guessing you don't have your customer number?	ElicitSlot	
C: Nah I don't have that what do I need it for?	InformIntent	
A: Huh it's fine I can just use some other infomation to complete the ID check.		HesitationsFillersLow, MissingInfoWorkaround
C: Whew alright.		
A: Could I get your date of birth$_{\text{DATEOFBIRTH}}$, phone number$_{\text{PHONENUMBER}}$ and social security$_{\text{SOCIALSECURITYNUMBER}}$?	ElicitSlot	MultiElicit
...		

Table 21: Example conversation excerpt from NATCS$_{\text{SELF}}$ demonstrating both spoken complexities as well as discourse complexities.

Text	Annotations	Complexities
A: Rivertown Insurance, this is Carla speaking how can I help you?	ElicitIntent	
C: Um hi could I create an account and enroll in a plan over the phone?	InformIntent (CreateAccount, EnrollInPlan)	HesitationsFillersHigh, MultiIntentUpfront
A: Mhm yes sir I can help you with that. Do you want to start the sign up now?	ElicitIntent	
C: Uh-huh can I only sign up over the phone?	InformIntent (AskFAQ)	FollowUpQuestion
A: Oh, no you could also use our website to create your account if you		
C: Eee I think I'll just have you help me .	InformIntent	HesitationsFillersHigh, Interruption
A: Sure thing sir first can I get your first and last name$_{\text{FIRSTNAME}}$last name$_{\text{LASTNAME}}$ please?	ElicitSlot	MultiElicit
C: Uh yeah Jonny$_{\text{FIRSTNAME}}$, j.o.n.n.y.	InformSlot	Spelling
A: Mhm.		Backchanneling
C: Barbados$_{\text{LASTNAME}}$ b.a.r.b.a.d.o.s.$_{\text{LASTNAME}}$	InformSlot	Spelling
A: Uh-huh thank you could I get your phone number$_{\text{PHONENUMBER}}$ now?	ElicitSlot	
C: Mhm five four three$_{\text{PHONENUMBER}}$.	InformSlot	
A: Five four three$_{\text{PHONENUMBER}}$.	ConfirmSlot	Confirmation
...		

Table 22: Example annotated conversation excerpt from NATCS$_{\text{SELF}}$ demonstrating both spoken complexities (e.g. Backchanneling) as well as discourse complexities (e.g. MultiIntentUpfront and FollowUpQuestion).

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
7

☑ A2. Did you discuss any potential risks of your work?
7

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

3,4,5

☑ B1. Did you cite the creators of artifacts you used?
2,3,4,5

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
License information is embedded in the released datasets.

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
The artifacts used (models) have standard licenses and expectations of use. The created artifacts are not derivative, and usage restrictions are embedded in the release license.

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
8 (Ethics Statement) and 4 (Dataset Analysis), discussing PII redaction.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
8 (Ethics Statement)

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
4 (Dataset Analysis)

C ☑ Did you run computational experiments?

4 (Dataset Analysis), 5 (Applications), Appendix

☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
The paper does not focus on modeling improvements, but instead of analysis of data – these details were deemed unnecessary.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
4 (Dataset Analysis), 5 (Applications)

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
4 (Dataset Analysis), 5 (Applications), Appendix

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

3 (Methodology), 4 (Dataset Analysis), 8 (Ethics Statement)

☒ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Partial instructions were provided (Appendix), however, full details cannot be released for legal reasons.

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
8 (Ethics Statement)

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
8 (Ethics Statement)

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not requires for this kind of data collection.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
8 (Ethics Statement)