DynDST: A Dynamic Dialogue State Tracking Dataset for Assessing the Conversational Adaptability of Large Language Models

Anonymous ACL submission

Abstract

This work tackles a key challenge in dialogue systems: the ability to adapt to changing user intentions and resolve inconsistencies in conversation histories. This is crucial in scenarios like train ticket booking, where customer plans often change dynamically. Despite advancements in NLP and large language models (LLMs), these systems struggle with real-time information updates during conversations. We introduce a specialized dataset to evaluate chatbot models on dynamic dialogue state tracking, focusing on scenarios where users modify their requests mid-conversation. This work aims to improve chatbot coherence and consistency, bridging the gap between the current capabilities of dialogue systems and the fluidity of human-like conversational interactions.

1 Introduction

In the dynamic flow of a conversation, it is common for speakers to shift their intentions and revise their previously spoken words. Take, for instance, the scenario of a customer booking train tickets for travel. It's often the case that the customer's initial travel plans are subject to change during the booking process, influenced by factors like ticket availability. In response to these changes, the booking agent, responsible for understanding and processing the customer's intent, must promptly update their comprehension of the customer's needs and adapt their responses to align with the customer's latest requirements.

As dialogue systems continue to evolve, an increasing number of online customer service interactions are being managed by NLP models. Yet, the ability of these models, including the most advanced large language models (LLMs), to accurately and efficiently update information during a conversation remains a significant challenge. This difficulty stems from the need for the chatbot model to not only understand the nuances of human communication but also to dynamically adjust its understanding on the fly as the conversation progresses and new information emerges.

The crux of the issue lies in the model's ability to discern and adapt to the latest user intent, effectively disregarding or re-contextualizing the outdated information from earlier in the conversation. This requires the model not only to understand the current request but also to identify and resolve the inconsistencies within the conversation history. Achieving this would enable the chatbot to respond accurately to the user's most recent requirements and intents.

The problem intensifies as conversation histories grow longer and changes become more frequent or subtle. The model must continuously track the conversation, identify shifts in context or intent, and reconcile any discrepancies in the information flow. This requires advanced capabilities in contextual understanding, memory management, and dynamic response generation, pushing the boundaries of current NLP technologies.

In essence, the ability of a chatbot to effectively manage and resolve inconsistencies in conversation histories, aligning itself with the user's latest intents and requirements, is pivotal in enhancing the efficiency and reliability of dialogue systems. As NLP models evolve, addressing this challenge will be crucial in bridging the gap between human-like conversational agility and the current capabilities of automated dialogue systems. Therefore, tackling consistency is a never-ending challenge in the development of dialogue systems (Vinyals and Le, 2015; Li et al., 2016; Zhang et al., 2018), and several training approaches have been proposed to enhance chatbot coherence (Yi et al., 2019; Li et al., 2020; Bao et al., 2021; Ouyang et al., 2022). To evaluate the consistency capacity, existing benchmarks for contradiction detection (Welleck et al., 2019; Nie et al., 2021; Zheng et al., 2022) treat contradictory responses from chatbots as errors.

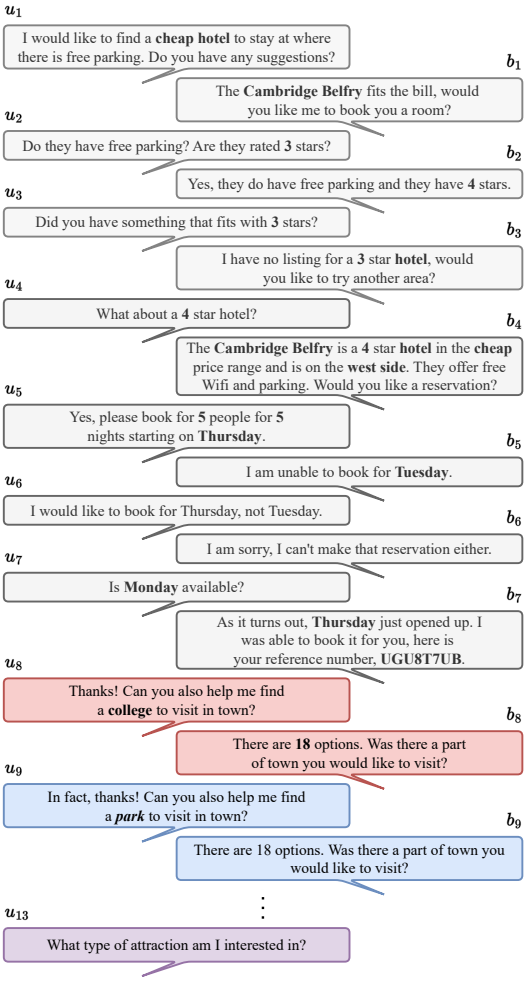

Figure 1: An example of our DynDST dataset. The customer made a wrong request in u_8 and then indicated the inquired type of attraction is park in u_9.

Note that existing dialogue contradiction datasets, in essence, can be reduced to the *bot* response b_i contradicting its previous response b_j. More importantly, they do not consider whether the information has been rendered obsolete or updated by the *user* either. This work presents a challenging dataset differs from those aforementioned datasets. Our dataset aims at evaluating the ability of chatbot models for dynamically tracking the user state in the dialogue. Figure 1 displays an instance of our dataset. The customer inquired a wrong attraction (*i.e.*, college) in u_8 and then immediately made an update in u_9. A reasonable chatbot model should adapt to this change and provide options of park instead of college.

In this paragraph, we briefly describe how to generate our DynDST dataset (details are in Section 3). We extend the MultiWOZ 2.2 dataset (Zang et al., 2020; Eric et al., 2020; Budzianowski et al., 2018) by first identifying all the slots or text span (highlighted in **bold**) in the dialogue, then we randomly choose one *user* utterance and alter one of its entities. As shown in Figure 1, college is selected and u_8, which, along with the corresponding bot response (b_8), is considered a false turn. Next, we duplicate the false turn and make any necessary changes obtain the correct turn (u_9 and b_9). Finally, we gather the question (u_{13}) inquiring whether the incorrect state has been overwritten. In this example, the model should output park (new answer) instead of college (old answer) in b_{13}.

Our contributions are three-fold:

- We delve into the challenging zero-shot, in-context knowledge editing task for LLMs, which we believe intelligent LLMs shall generate responses that are not only consistent but also *adaptive* in long-term conversations.[1]
- We construct the DynDST dataset that serves as a benchmark for the evaluation of chatbot's adaptability. The dataset has 8,001 examples.
- We propose a parameter-free method to perform "exact match" criterion for generative models, which is sometimes problematic as such models are uncontrollable (not to mention most of the LLMs are not accessible and can only inference through their APIs). We show our approach is effective in the preliminary experiment and it alleviates the need of prompt engineering and is applicable to open-domain question.

2 Definition

Fact The term fact refers to the text to be edited throughout this paper. Seeing that the MultiWOZ dataset solely focuses on tracking the personal status (*e.g.*, booking a hotel), they are not intrinsically pertain to the *factual* knowledge in the real-world. In other works, it may have different definitions, names, and even forms (Mitchell et al., 2022b; Meng et al., 2023). We follow the form of fact in Meng et al. (2023), which is a tuple τ comprising subject, relation, and object. Intuitively, given a fact x, we define the new fact x' is semantically different (*i.e.*, x' is *effective*) as:

$$\tau(x') \neq \tau(x) \quad (1)$$

[1]Here, "in-context" is different from Brown et al. (2020).

Dataset	Lang	Factual? F	Factual? ¬ F	Eff	LT	# Turn (m, M)	Source
zsRE	en	✓	✗	✗	–	–	Levy et al. (2017)
FEVER	en	✓	✗	✓	–	–	Thorne et al. (2018)
Dialogue NLI	en	✗	✓	✓	–	–	Welleck et al. (2019)
COUNTERFACT	en	✓	✗	✓	–	–	Meng et al. (2022)
TruthfulQA	en	✓	✗	✓	–	–	Lin et al. (2022)
Wikitext generation	en	✓	✗	✓	–	–	Mitchell et al. (2022a)
DECODE	en	✗	✓	✓	✗	(4.4, 4.5)	Nie et al. (2021)
CareCall$_{mem}$	ko	✗	✓	✓	✓	(12.0, 11.5)†	Bae et al. (2022)
DIALFACT	en	✓	✗	✓	✗	(2.7, 2.5)	Gupta et al. (2022)
CDCONV	zh	✗	✓	✓	✗	(2.0, 2.0)	Zheng et al. (2022)
DynDST (Our)	en	✗	✓	✓	✓	(7.9, 8.0)	

† We report the English version

Table 1: An overview of various datasets from their *source* papers. The data attributes and statistics presented are exclusively pertain to the **contradiction** relation in NLI or knowledge editing. In this table, we separated these datasets from their original input format; the upper half is either in sentence or paragraph (substantially longer sequence) format, while the the lower half is in chat format. Lang stands for language. F/¬ F column displays if the dataset contains factual/non-factual knowledge to be edited. Eff stands for effective (defined in Section 2). LT stands for long-term. Though there is no definite number of long-term, we regard the dataset as long-term so long as half of the data have at least 5 turns (note that a conversation turn is defined as a pair of user and chatbot utterances. The underlined checkmark (✓) denotes the source data that partially satisfies the property. We also report the mean (m) and median (M) number of turns in the # Turn column, if the input data can be converted to the chat format.

Conversation A conversation or dialogue with n turns is denoted as $(u_1, b_1, ..., u_n, b_n)$, where u_i and b_i is the user and bot utterance in the i-th turn, respectively. We focus on whether b_{n+1} updates the fact in the dialogue context when a question related to such fact is asked in u_{i+1}, given an effective fact introduced in the user utterances. We decompose a **multi-turn** conversation into four disjoint turns; namely, *false, update, test, and previous turn*. Simply put, (1) the false turn contains a false fact; (2) the update turn has user utterance that corrects the previous false turn; (3) the test turn is the question we aim to assess whether the chatbot pays attention to the user correction in the update turn; and (4) the rest of turns fall into the previous turn.

3 Experimental Setup

Dataset Generation As our main goal is to generate a dataset that user updates their status, we first filter out data that does not have any labeled text span in user utterance in MultiWOZ 2.2 training dataset. After setting random seed to 0, we randomly select one utterance for each data and obtain the first slot to be edited. To generate another slot that is semantically different, we gather the (universal) set from all training data, then we randomly select one element that does not include in the current data. Mathematically speaking, let $\mathcal{D} = \{d_1, d_2, ...\}$ be the training set, $\mathcal{U}(\mathcal{D}) = \bigcup_i \mathcal{U}(d_i)$ be the set union of slots on all data in $\mathcal{D}$. For each d_i and its associated slot s_i to be edited, another effective slot s' is picked from $\mathcal{U}(\mathcal{D}) \setminus \mathcal{U}(d_i)$, where s' has the same slot name as s_i (*e.g.*, *hotel-pricerange*, *train-bookpeople*). The final number of training data in the DynDST dataset is 8,001.

Model In order to evaluate the most recent LLM's adaptability in a multi-turn fashion, we utilize `gpt-3.5-turbo-0125` version of GPT-3.5 and `gpt-4-0125-preview` version of GPT-4. To stabilize the performance, we run our DynDST dataset five times. The top_p, frequency_penalty, presence_penalty, and temperature is set to 1, 0, 0, and 0 to maximize the reproducibility.

Framework Note that the location of the update turn, whether it is more contextualized to the false turn or the test turn, also largely affect the result in our pilot study, and we choose the scenario that users *immediately* correct themselves in this paper. **Exp. 1** is the baseline, where we test the original

	Update (↑, Maj)			No Update (↓, Maj)			Oracle (↑)		
Top-K	1	3	5	1	3	5	1	3	5
Exp. 1 (baseline)	66.6	72.5	72.6	20.7	19.7	20.7	66.6	80.7	83.7
Exp. 1 (baseline) [GPT-4]	58.4	64.3	66.5	24.5	24.8	26.3	58.4	76.1	80.3
(a) w/o choice	62.6	66.6	69.4	20.3	21.7	21.9	62.6	76.1	79.8
(b) w/o long	46.7	51.1	50.5	39.7	39.9	42.6	46.7	67.8	73.7
(c) w/o both	42.3	48.3	47.2	39.8	40.0	43.6	42.3	61.8	67.2
Exp. 2 (Our)	73.7	79.9	**80.2**	13.9	13.4	14.7	73.7	86.3	88.7
(a) w/o choice	70.3	75.1	78.7	16.5	15.6	14.5	70.3	82.4	86.4
(b) w/o long	70.1	75.9	76.9	16.1	16.0	17.2	70.9	84.1	88.3
(c) w/o both	68.8	74.2	76.6	13.9	15.8	16.2	68.8	81.7	86.4

Table 2: Percentage of Update/No Update on DynDST dataset. Maj stands for majority voting. Oracle column represents an upper bound performance in which a successful update occurs if *any* run triggers the model to respond accordingly among K templates based on the correction turn. All results are reported using GPT-3.5 except the second row in **Exp. 1**, where we utilize GPT-4. The sum of Update and No Update is not 100, as we exclude invalid response in the table; this also happens if there is a tie in majority voting. We also report the ablation analysis of two experiments with the removal of (a) choice in test turn, (b) long correction in update turn, and (c) both.

DynDST dataset. In **Exp. 2**, we inject some *predefined* sequences into the data in the hope of eliciting the correct answer of LLMs in the test turn; the string prepended to the bot utterance in correct turn is "No problem at all! I have updated my memory with the correction you provided. Thank you for letting me know." The user utterance also contain context that explicitly negate the false statement. For instance, one template used in our experiment is "I'm sorry to bring this up, but I mistakenly gave you `[X]`. In fact, `[Y]`," where `[X]` and `[Y]` are the slots for the false and correct user utterance.

Evaluation Metric In general, we employ an exact match criterion by extracting and comparing the LLM output and the gold answer, which is widely used in knowledge editing (De Cao et al., 2021; Mitchell et al., 2022b; Meng et al., 2023) with some twist lest we *underestimate* the LLM's capability. We briefly state how we combine the exact match (EM), ROGUE-1 (R-1), and ROGUE-L (R-L). First, we seek to EM defined the process as success (failure) if the new (old) answer is *exclusively* in the model output. Next, we convert data to its "canonical" form and perform EM again. After that, we remove the punctuation and stop words (with the aid of NLTK package and our automatically method that can generate GPT-3.5's own stop words set), and execute EM again. At last, we use the *strict* rule to compute the R-1 and R-L. We compute R-1/R-L score of old answer and new answer with the model output. In R-1, the model output is considered new only if the F1 is larger and either its precision or recall is larger than $\max\{0.5, \text{old}\}$. In R-L, we combine the edit distance and longest common substring.

4 Results and Discussion

The results are tabulated in Table 2. The choice in this table means we provide the model some hints after we question the model at the test turn. The short correction (*i.e.*, w/o long) is we only fill the templates with the text span instead of the entire utterance. Our results demonstrate that when selecting the top 5 templates and making decisions through majority voting, GPT-3.5, on average, tends to update the knowledge by more than 70% in Exp. 1 and slightly above 80% in the Exp. 2.

Note that Exp. 2 consistently outperforms Exp. 1 across all settings, indicating that the injected sequence in bot utterance will boost GPT-3.5 to pay more attention to correction turn. Moreover, the table shows that Exp. 2 still outperforms Exp. 1 even if they are in setting (c). Lastly, we point out that while there is a common belief that GPT-3.5 is bested by GPT-4 in every tasks, GPT-3.5 significantly outperform GPT-4 in our dataset.

5 Conclusion

Unlike existing DST datasets that primarily assess whether chatbots could incrementally expand the state of single domain or perform multiple tasks in the same dialogue, we construct our DynDST dataset to evaluate the model's capacity for recognizing and deleting state in long-term conversations. We hope our work will inspire future research to build a better chatbot for long-term companion.

Limitations

When evaluating the results, our exact match method, though demonstrate it can catch nuances of typos, is not flawless, so it may produce unwanted results, even if we have experimented adding another constraint: the edit distance (ED), longest common subsequence (LCSeq), and longest common substring (LCStr) due to numerous typos (similarly, the model output is considered new only if new's ED $<$ old's ED $\wedge$ new's LCSeq $>$ old's LCSeq $\wedge$ new's LCStr $>$ half of the new's string length). For instance, suppose the slot name is *restaurant-food*, the old answer is "North Indian", and the new answer is "Labanese" (typo, should be "Lebanese") in the dataset. If model outputs "Japanese", our evaluation will consider it correct in Step 5. Moreover, it may require thousands of related data so that we can generate the model's own stop words set. This paper is the pioneer study on deleting existing state in dialogue state tracking task, so the experiments do not cover a variety of open-domain LLMs. Consequently, testing whether other LLM-based chatbots are on par with state-of-the-art GPT models is also a promising avenue of research. Likewise, there is potential for future research to explore better templates in our pre-defined texts. For example, we can provide the model with few-shot examples in each phase or methodology; however, note that the selection of best examples and the order of selected demonstrations within context may require extensive experiments to meet the needs (Zhao et al., 2021).

Ethical Statement

It is important to note that the LLM should not be treated as an authoritative source of facts, although we test the LLM's adaptability and treat its output as the definite answer. As the DynDST dataset is constructed based on the MultiWOZ 2.2, we do not foresee any ethical issues in the dataset.

References

Sanghwan Bae, Donghyun Kwak, Soyoung Kang, Min Young Lee, Sungdong Kim, Yuin Jeong, Hyeri Kim, Sang-Woo Lee, Woomyoung Park, and Nako Sung. 2022. Keep me updated! memory management in long-term conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3769–3787, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Siqi Bao, Huang He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhen Guo, Zhibin Liu, and Xinchao Xu. 2021. PLATO-2: Towards building an open-domain chatbot via curriculum learning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2513–2525, Online. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

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.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6491–6506, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. 2020. MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 422–428, Marseille, France. European Language Resources Association.

Prakhar Gupta, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. DialFact: A benchmark for fact-checking in dialogue. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3785–3801, Dublin, Ireland. Association for Computational Linguistics.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342, Vancouver, Canada. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long*

Papers), pages 994–1003, Berlin, Germany. Association for Computational Linguistics.

Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. 2020. Don't say that! making inconsistent dialogue unlikely with unlikelihood training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4715–4728, Online. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372. Curran Associates, Inc.

Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations*.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. 2022a. Fast model editing at scale. In *International Conference on Learning Representations*.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D Manning, and Chelsea Finn. 2022b. Memory-based model editing at scale. In *International Conference on Machine Learning*, pages 15817–15831. PMLR.

Yixin Nie, Mary Williamson, Mohit Bansal, Douwe Kiela, and Jason Weston. 2021. I like fish, especially dolphins: Addressing contradictions in dialogue modeling. 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 1699–1713, Online. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741, Florence, Italy. Association for Computational Linguistics.

Sanghyun Yi, Rahul Goel, Chandra Khatri, Alessandra Cervone, Tagyoung Chung, Behnam Hedayatnia, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tur. 2019. Towards coherent and engaging spoken dialog response generation using automatic conversation evaluators. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 65–75, Tokyo, Japan. Association for Computational Linguistics.

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. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 109–117, Online. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

Chujie Zheng, Jinfeng Zhou, Yinhe Zheng, Libiao Peng, Zhen Guo, Wenquan Wu, Zheng-Yu Niu, Hua Wu, and Minlie Huang. 2022. CDConv: A benchmark for contradiction detection in Chinese conversations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 18–29, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.