SYNDICOM: Improving Conversational Commonsense with Error-Injection and Natural Language Feedback

Anonymous ACL submission

Abstract

Commonsense reasoning is a critical aspect of human communication. Despite recent advances in conversational AI driven by large language models, commonsense reasoning remains a challenging task. In this work, we introduce SYNDICOM - a method for improving commonsense in dialogue response generation. SYNDICOM consists of two components. The first component is a dataset composed of commonsense dialogues created from a knowledge graph and synthesized into natural language. This dataset includes both valid and invalid responses to dialogue contexts, along with natural language feedback (NLF) for the invalid responses. The second contribution is a two-step procedure: training a model to predict natural language feedback (NLF) for invalid responses, and then training a response generation model conditioned on the predicted NLF, the invalid response, and the dialogue.

SYNDICOM is scalable and does not require reinforcement learning. Empirical results on three tasks are evaluated using a broad range of metrics. SYNDICOM achieves a relative improvement of 53% over ChatGPT on ROUGE-1, and human evaluators prefer SYNDICOM over ChatGPT 57% of the time. We will publicly release the code and the full dataset.

1 Introduction

Conversational AI has witnessed rapid advancements in recent years, largely due to the success of large language models (LLMs) such as GPT-3 (Brown et al., 2020). These advancements have been driven by the notable achievements of models like ChatGPT, which is built upon InstructGPT (Ouyang et al., 2022). InstructGPT was trained on an extensive dataset of instructions for various language tasks and was further enhanced using human feedback and reinforcement learning (RL). Consequently, research in conversational AI has shifted towards leveraging large models trained on extensive datasets, supplemented by human feedback.

While these models have consistently demonstrated significant improvements in reasoning and problem-solving capabilities, they still exhibit flaws and issues. In many critical applications of LLMs, the tolerance for errors in dialogue responses is exceedingly low. Addressing these problems remains challenging, primarily due to the scarcity of data and the high cost associated with human feedback. Recent research has started exploring alternative techniques beyond human feedback and RL, such as natural language feedback (NLF) and self-correction (Saunders et al., 2022; Scheurer et al., 2022; Welleck et al., 2022; Bai et al., 2022b).

Furthermore, even with the progress made, large models often generate hallucinations, underscoring the ongoing importance of knowledge grounding. One of the most demanding aspects of knowledge grounding is commonsense knowledge. Recent advancements in incorporating commonsense into LLMs have utilized resources such as ConceptNet (Speer et al., 2017) or ATOMIC (Sap et al., 2019).

This paper presents a method for improving commonsense dialogue responses by (1) replacing human feedback and RL with natural language responses and (2) leveraging recent knowledge graph techniques to ground responses in commonsense knowledge derived from ATOMIC. To address the scarcity of data and the high cost of human feedback, the natural language feedback is elicited in a manner that specifically targets the chosen error types determined by the designer. This approach significantly enhances the speed and quality of model learning and refinement.

The contributions of this paper are as follows:

- Development of a scalable method for synthesizing knowledge-grounded data with error injection and feedback.

- Release of a dataset rich in dialogues featuring commonsense inferences, annotated with

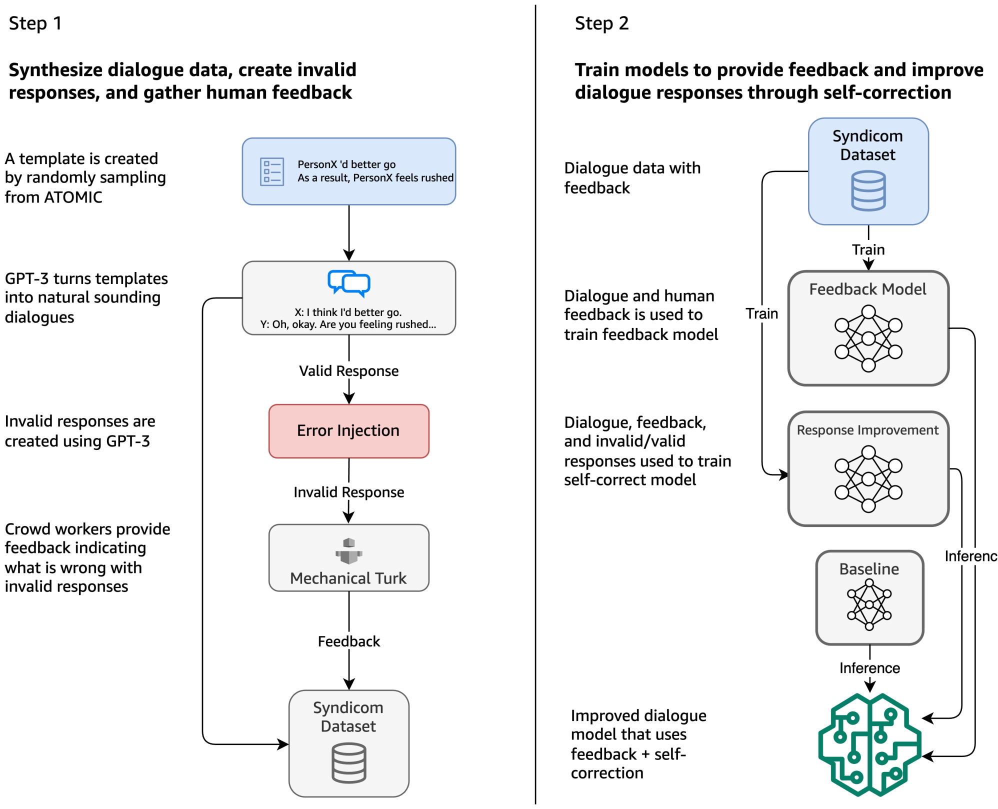

Step 1
Synthesize dialogue data, create invalid responses, and gather human feedback
A template is created by randomly sampling from ATOMIC
PersonX 'd better go
As a result, PersonX feels rushed
GPT-3 turns templates into natural sounding dialogues
X: I think I'd better go.
Y: Oh, okay. Are you feeling rushed...
Valid Response
Invalid responses are created using GPT-3
Error Injection
Invalid Response
Crowd workers provide feedback indicating what is wrong with invalid responses
Mechanical Turk
Feedback
Syndicom Dataset
Step 2
Train models to provide feedback and improve dialogue responses through self-correction
Dialogue data with feedback
Syndicom Dataset
Train
Dialogue and human feedback is used to train feedback model
Train
Feedback Model
Dialogue, feedback, and invalid/valid responses used to train self-correct model
Response Improvement
Inference
Baseline
Inference
Improved dialogue model that uses feedback + self-correction

Figure 1: SYNDICOM Process. Left: dataset generation, Right: Improving commonsense in dialogue response generation.

commonsense errors, and accompanied by human-written feedback, which we refer to as SYNDICOM.

- Description of a method for training both a feedback generation model and a response improvement model using natural language feedback (NLF), and demonstration of the superiority of this information-rich approach over state-of-the-art RL methods using SYNDICOM.

2 Recent Work

The field of conversational AI has experienced a surge of interest in commonsense reasoning in recent years, with a significant focus on curating datasets (Richardson and Heck, 2023). ConceptNet (Speer et al., 2017) and ATOMIC (Sap et al., 2019) have emerged as widely used resources for dataset curation, establishing a de facto standard. Several datasets serve as sources for the dialogues, including DailyDialogue (Li et al., 2017), MuTual (Cui et al., 2020), DREAM (Sun et al., 2019), and the Ubuntu Dialogue Corpus (Lowe et al., 2015).

Our research lies at the intersection of two critical areas in conversational AI: the synthesis of commonsense datasets and the training of models using natural language feedback. These areas have recently garnered significant research attention due to their potential to enhance the ability of conversational agents to understand and respond to complex human interactions with greater accuracy and consistency. By leveraging the synergies between these domains, our work aims to address the existing limitations in conversational agents and pave the way for more robust and effective conversational systems.

2.1 Commonsense Dataset Curation

In recent years, various datasets have been curated specifically for commonsense reasoning. Ghosal et al. (2021) introduced CIDER, a dialogue dataset annotated with commonsense inferences, which was later expanded with the more open-ended CICERO (Ghosal et al., 2022). Some researchers have focused on specific types of commonsense, such as temporal commonsense (Qin et al., 2021) and ethical commonsense (Ziems et al., 2022; Kim et al., 2022; Sun et al., 2022). Others have concentrated on grounding dialogues in knowledge graphs

(Zhou et al., 2021a; Moon et al., 2019).

These approaches rely on existing dialogue datasets and often employ filtering strategies to reduce dataset size. However, this reliance on existing datasets can limit the generalizability of methods to future problems. One potential solution to the scarcity of large-scale annotated commonsense knowledge datasets is the synthesis approach. Recently, Kim et al. (2022) proposed SODA, a method for procedurally generating social dialogues based on a commonsense knowledge graph. They utilized ATOMIC (Sap et al., 2019), which consists of atomic facts in natural language form, to generate synthetic dialogues rich in commonsense inferences. Their entirely procedural and highly scalable approach generates dialogue data suitable for training models that reason over commonsense knowledge. Building upon this work, we present SYNDICOM, a synthesis procedure and dataset that expands on the ideas of SODA and incorporates novel features crucial for our dialogue modeling approach. More details about SYNDICOM are provided in Section 3.

2.2 Feedback and Response Improvement

The use of feedback to improve language models has recently garnered increased interest, with most efforts focused on the application of reinforcement learning (Stiennon et al., 2020; Zhou et al., 2021b; Bai et al., 2022a,b). Reinforcement learning with human feedback (RLHF) is particularly notable as it serves as the foundation for InstructGPT (Ouyang et al., 2022), which paved the way for ChatGPT. RLHF offers a flexible approach to improving LLMs; however, it faces challenges in terms of stability and efficiency inherent to RL. Moreover, the low dimensionality of the reward signal in RL (typically a scalar) severely limits the learning rate.

A more information-rich approach than RL is the use of natural language feedback (NLF). NLF has been explored in several recent works. Scheurer et al. (2022) investigated the use of human-written NLF to train a dialogue response refinement model. Saunders et al. (2022) demonstrated that LLMs themselves can generate this feedback. Welleck et al. (2022) developed a method to improve sequence generation of LLMs by first generating a baseline using an imperfect base generator and then correcting the output using a second correction model. The correction model incorporates feedback as part of its input. However, the authors only demonstrated the use of feedback provided by various tools and APIs tailored to the specific tasks they explored.

3 The SYNDICOM Method

Taking inspiration from recent NLF methods, this paper presents a new approach called SYNDICOM. This new approach combines the synthesis of commonsense dialogue data from a grounded knowledge graph (ATOMIC) with an NLF response improvement approach to improve dialogue responses. Figure 1 illustrates the two phase process.

3.1 SYNDICOM Dataset

The SYNDICOM dataset is created in a four step process: (1) Auto-generate commonsense dialogue templates, (2) Translate templates into natural language dialogues, (3) Generate invalid responses with error injection, and (4) Collect human-written explanations for the invalid responses. Examples from the SYNDICOM dataset are shown in Table 1. The GPT model we used for the steps in this section was text−davinci−003.

3.1.1 Generating Templates

Our approach generates commonsense-focused dialogue templates from a commonsense knowledge base. For this study, we utilize ATOMIC (Hwang et al., 2021). ATOMIC consists of inferences in the form of Head $\xrightarrow{\text{relation}}$ Tail. Each head and tail is a natural language description of a generic event, emotional state, action, description, etc. Dialogue templates are constructed by crawling through inferences rooted at each head of ATOMIC and chaining these inferences together to form multiple dialogue turns. The number of dialogue template turns is uniformly and randomly chosen between 3 and 8.

3.1.2 Converting to Natural Language

Given the dialogue templates, the second step in creating SYNDICOM converts the templates to natural language conversations. We explored several methods, including crowdsourcing, but found LLMs to be the most consistent and effective. We used the GPT LLM (text-davinci-003) to generate the natural language dialogues from the templates. This was followed by in-context learning with 15 hand-written examples. The exact prompting used is shown in detail in Appendix A.

Template	Synthesized Dialogue	Explanation
PersonX refuses PersonY ↪ PersonX is seen as: disagreeable ↪ As a result, PersonX feels: annoyed and irritated ↪ Before that, PersonX needed: thinks about it	**Context** I refuse to do what you ask. *Why are you being so disagreeable?* I'm just annoyed and irritated. **Valid Response:** You should think about it before you say no. **Invalid Response:** You should think about it before you say yes.	**Crowd Worker 1:** The other person already said no. **Crowd Worker 2:** The person did not say yes so this response was strange.
PersonX makes music ↪ As a result, PersonX wants: to impress ↪ PersonX is seen as: talented ↪ As a result, PersonX will: gets asked to play something ↪ Before that, PersonX needed: to carry their violin ↪ PersonX wanted: they want to share their creativity	**Context** I love making music. *You're very talented!* Thanks. I just want to impress people with my playing. *I'm sure you will. Can you play something for me?* Of course! I always carry my violin with me. **Valid Response:** That's great. I'm glad you want to share your creativity. **Invalid Response:** That's awful. I don't want to share my creativity.	**Crowd Worker 1:** This contradicts with what was said about impressing people. **Crowd Worker 2:** They aren't being asked to play. They asked the other person to play

Table 1: Example dialogues from SYNDICOM. Each dialogue context includes both valid and invalid responses, as well as crowd worker-written explanations for the invalid response (italicized).

3.1.3 Error Injection

To elicit feedback on commonsense from crowd workers, the SYNDICOM process starts by corrupting the valid dialogue responses so that they violate commonsense reasoning. This provides crowd workers with an easy target for their feedback. To corrupt the dialogue responses, SYNDICOM takes advantage of the commonsense dialogue inference structure provided by ATOMIC. Given a commonsense knowledge base $\mathcal{K}$, a dialogue context $\mathcal{C}$, and response r from SYNDICOM, the response is implied by commonsense from the context, or $\mathcal{C} \xrightarrow{\mathcal{K}} r$. The response r is corrupted by replacing it with the semantic opposite, $\overline{r}$. We prompted GPT as shown in Appendix A to acquire these semantic opposites. The result is dialogues annotated with commonsense contradictions of the form $\{\mathcal{C}, r, \overline{r}\}$.

3.1.4 Natural Language Feedback Acquisition

The dialogues with commonsense contradictions are presented to crowd workers on the Amazon's Mechanical Turk platform. Each dialogue is shown in the form of context and invalid responses, informing them that the dialogues were generated by an AI attempting to sound human. The crowd workers were given instructions to review AI-generated casual text message conversations and provide 1-2 sentences of natural language feedback on the dialogue, and the final turn in particular (the invalid response). They were asked to be as specific as possible in their feedback. The full instructions and web interface given to the crowd workers can be found in Appendix A.

To ensure the quality of the feedback, we used only masters-level crowd workers from English-speaking countries. This decision aimed to maximize the clarity and accuracy of the feedback provided. Each dialogue was evaluated by two crowd workers independently, allowing for a more comprehensive understanding of the AI's mistakes and ensuring a diverse range of feedback.

With the addition of the feedback f, this completes the dataset synthesis part, resulting in annotated dialogues of the form $\{\mathcal{C}, r, f, \overline{r}\}$.

3.2 SYNDICOM Dialogue Improvement

This section details the process of using natural language feedback to correct latent errors in the baseline conversational response. To begin, the dialogue response improvement problem is defined as follows: given a dialogue context $\mathcal{D}$ and a response r_b, generated by some dialogue system or model, produce an improved response r^*.

$$r^* = \underset{r}{\operatorname{argmax}}\, p(r|\mathcal{D}, r_b) \tag{1}$$

Dialogue response generation and improvement has recently received considerable attention (Shah et al., 2016; Nayak et al., 2017; Liu et al., 2017, 2018; Weston et al., 2018). This problem is especially relevant today with large language models (LLMs). While LLMs have recently reached a high degree of fluency in dialogue, in some domains they can be factually inaccurate. While these cases are relatively infrequent, the tolerance for factual errors for a number of important applications is very low. In addition, these errors are difficult to predict and/or automatically detect. This leads to a problem of data sparsity that is difficult to overcome for response improvement methods that rely on training models.

Hyperparameter	Value
Temperature	0.7
Max tokens	50
Top p	1.0
Frequency penalty	0
Presence penalty	0

Table 2: Hyperparameters used for GPT-3.5. The same parameters were used for training and inference.

A method to partially mitigate the sparsity of dialogue response errors is to *artificially create invalid responses* $\overline{r}$ via error injection (as described in Section 3.1.3). This method will be called Syndicom-Direct. Given the invalid response $\overline{r}$ and the dialogue history $\mathcal{D}$, a model is trained to learn the optimal response r^*

$$r^* = \underset{r}{\operatorname{argmax}}\, p(r|\mathcal{D}, \overline{r}). \tag{2}$$

A second approach called Syndicom-NLHF includes natural language human feedback (NLHF) to explain the rationale for why the response $\overline{r}$ is invalid and then conditions on this side rationale.

$$r^* = \underset{r}{\operatorname{argmax}}\, p(r|\mathcal{D}, \overline{r}, f^*). \tag{3}$$

As a comparison, we also implemented an approach called Syndicom-Multistep. This approach breaks the inclusion of NLHF into two steps: (1) train a feedback model on NLHF that *predicts* the feedback critical of response $\overline{r}$

$$\hat{f} = \underset{f}{\operatorname{argmax}}\, p(f|\mathcal{D}, \overline{r}). \tag{4}$$

and (2) train a second model to produce an improved dialogue response from the invalid response, given the *predicted* feedback

$$r^* = \underset{r}{\operatorname{argmax}}\, p(r|\mathcal{D}, \overline{r}, \hat{f}). \tag{5}$$

Both models used in this work are based on OpenAI's GPT-3.5, specifically text−davinci−003. The models were fine-tuned through the OpenAI API for GPT based models. The hyperparameters used are listed in Table 2.

4 Experiments

In this section, we provide a detailed description of the experiments conducted to evaluate our proposed method, Syndicom. The experiments aim to compare the direct prediction of the improved response in Equation 2 (Syndicom-Direct) with the response prediction when conditioned on natural language human feedback (NLHF) that explains why the initial response is invalid (Syndicom-NLHF). Additionally, we explore a multistep implementation of NLHF (Syndicom-Multistep). We compare the performance of our method against a ChatGPT baseline (gpt−3.5−turbo) using various text generation metrics, such as ROUGE, BLEU, SacreBLEU, BERTScore, and METEOR.

4.1 Syndicom-Direct

Our first experiment focused on the direct dialogue improvement task, where the objective is to enhance a dialogue response based solely on the context and an invalid response. No feedback, whether human or generated, was involved in this task. This optimization problem is described in Equation 2.

In order to prevent the model from simply learning to undo the error injection, we introduced noise by rephrasing the invalid dialogues using an independent ChatGPT instance. This rephrasing was only performed at inference time and not during training. The rephrasing prompt is available in Appendix A.

4.2 Syndicom-Multistep

Next, we explored the Syndicom-Multistep approach. As shown in Equations 4 and 5, we first predicted feedback using the feedback model and then improved the dialogue response using the response improvement model. For the feedback predictor, we trained a GPT-based model to generate feedback given a dialogue context and an invalid response, as shown in Equation 4, using the typical causal language modeling objective. We evaluated the feedback generation model portion of Syndicom-Multistep separately and compared it to ChatGPT. The prompt used for the baseline can be found in Appendix A. Table 3 presents the results, demonstrating that our method outperformed the baseline on all metrics.

Subsequently, we utilized the predicted feedback along with the dialogue context and invalid response to produce an improved dialogue response, as shown in Equation 5. Similar to the Syndicom-Direct experiments, we applied rephrasing to the invalid responses at inference time. The baseline model was explicitly instructed to first generate feedback for the invalid response and then use that

	ChatGPT			SYNDICOM		
Metric	**Max**	**Min**	**Avg**	**Max**	**Min**	**Avg**
ROUGE1	0.204	0.123	0.163	0.315	0.185	0.250
ROUGE2	0.034	0.0078	0.0209	0.112	0.035	0.073
ROUGEL	0.150	0.093	0.122	0.248	0.144	0.196
BERTSCORE	0.863	0.853	0.858	0.883	0.866	0.874
SacreBLEU	2.546	1.533	2.039	6.697	2.907	4.802
BLEU	0.004	0.0001	0.0021	0.030	0.0041	0.0171
METEOR	0.197	0.129	0.163	0.279	0.158	0.219

Table 3: Performance in Feedback Generation performance of our method vs. baseline. SYNDICOM outperforms the baseline on all metrics. Each dialogue was accompanied by two feedback responses, and scores were computed for both independently. We show the max/min/avg over the two for each score and model.

feedback to guide its response improvement. Table 4 displays the results.

4.3 SYNDICOM-NLHF

The next experiment focused on enhancing dialogue responses using human feedback (Equation 3). Given a dialogue context, an invalid response, and human feedback, the goal was to generate an improved (valid) dialogue response. For this experiment, we utilized the raw human-written feedback from SYNDICOM and trained a separate GPT improvement model to generate valid responses. As before, we applied inference-time rephrasing to the invalid responses. Results are presented in Table 4 under SYNDICOM-NLHF. This version of our method outperformed the others on all metrics.

4.4 Human Evaluation

In addition to our automated metric evaluations, we conducted a human evaluation to assess the effectiveness of response improvements through generated feedback. This evaluation process mirrored the dialogue enhancement steps employed in the experiment described in Section 3.2.

It is important to note that task assignments for crowdworkers require explicit and precise definitions, which often pose challenges in evaluating the commonsense aspect through human intervention. Existing human evaluations primarily focus on assessing the accuracy of information or determining the most preferred output from a set of alternatives.

With the emergence of advanced language models like ChatGPT, human evaluation has become increasingly complex. This complexity arises from the remarkably high-quality and naturally articulated outputs generated by state-of-the-art models such as ChatGPT.

In our study, we instructed crowdworkers that an AI system was attempting to emulate human conversation and generate dialogue responses that align with commonsense understanding and fit the given context. The workers were presented with two distinct responses: a standard ChatGPT response and our SYNDICOM response. Their task was to select the response that appeared more human-like and natural.

Despite the impressive contextual relevance exhibited by ChatGPT responses, our method generated the more favored response **56.5%** of the time, compared to ChatGPT's 43.5% preference rate. For further details on the interface provided to the crowdworkers, please refer to Appendix A.

5 Discussion

In the Discussion section, we analyze the performance of our proposed SYNDICOM method in conversational AI compared to the baseline model ChatGPT. The results are summarized in Tables 3 and 4, where we observe that SYNDICOM outperforms ChatGPT on all automatic metrics for the feedback and dialogue response improvement tasks.

Specifically, Table 4 provides a comparison between our direct and multi-step approaches to the response improvement problem. Our multi-step method outperforms the direct method on various metrics such as ROUGE-1, BLEU, SacreBLEU, and BERTScore, despite the simplicity of the error typology used in the error injection during these experiments. This indicates that the multi-step approach has the potential to achieve even better performance when faced with more diverse error ty-

	ChatGPT		**SYNDICOM**		
Metric	Direct	NLHF	Direct	Multistep	NLHF
ROUGE1	0.132	0.231	0.386	**0.388**	*0.474*
ROUGE2	0.029	0.081	**0.174**	0.172	*0.246*
ROUGEL	0.112	0.201	**0.324**	0.322	*0.396*
BLEU	0.008	0.031	0.117	**0.125**	*0.168*
METEOR	0.209	0.290	**0.390**	0.387	*0.445*
SacreBLEU	0.885	3.107	11.716	**12.547**	*16.831*
BERTScore	0.859	0.880	0.909	**0.910**	*0.919*

Table 4: Response Improvement comparing ChatGPT with our new SYNDICOM methods. ChatGPT-Direct is fine-tuned to produce a valid response given only the invalid response, with no intermediate steps or feedback. ChatGPT-NLHF is additionally conditioned on natural language human feedback (NLHF). SYNDICOM-DIRECT is the model that optimizes Equation 2, SYNDICOM-MULTISTEP optimizes Equation 5, and SYNDICOM-NLHF conditions on the same NLHF as used by the ChatGPT models. Bold text illustrates the highest score between all methods that are not give NLHF, and italics indicate the highest scores among NLHF tasks. SYNDICOM outperforms the baseline on all metrics for both tasks.

pologies, which we leave as an avenue for future research.

One contributing factor to the superior performance of the multi-step method is the additional information encoded in the feedback model. The feedback model is trained on human feedback, providing it with more contextual information compared to the direct model, which is solely trained on valid and invalid responses. Even in cases where the direct model achieves slightly higher scores in certain metrics, the differences are negligible. Notably, BERTScore, which represents the most comprehensive model-based metric utilized in our evaluation, further supports the argument in favor of the multi-step approach with feedback generation.

When examining the NLHF columns in Table 4, we observe that SYNDICOM demonstrates significant improvement over ChatGPT for the response improvement task when provided with human feedback for the invalid response. This scenario aligns with use cases where feedback can be collected for a dialogue system and subsequently used to fine-tune and enhance the dialogue model. These findings underscore the value of the SYNDICOM method in continuous learning scenarios, particularly those where feedback from end users is actively being collected.

Overall, SYNDICOM exhibits strong performance compared to the state-of-the-art large language model ChatGPT, despite both models being based on the same underlying architecture (GPT-3.5). It is worth noting that ChatGPT underwent substantial reinforcement learning through human feedback during its refinement process, making the success of SYNDICOM even more noteworthy.

6 Conclusion

In this paper, we introduced SYNDICOM, a novel method for enhancing commonsense reasoning in dialogue response generation. By integrating a commonsense dialogue synthesis approach with targeted error injection, we tackled the challenge of incorporating commonsense knowledge into conversational AI systems. Our method comprised two key components: (1) a dataset consisting of valid and invalid responses to dialogue contexts, along with natural language feedback (NLF) for the invalid responses, and (2) a two-step procedure involving training a model to predict NLF for invalid responses, followed by training a response generation model conditioned on the predicted NLF, the invalid response, and the dialogue.

A notable advantage of SYNDICOM is its scalability and independence from reinforcement learning techniques, which are commonly employed in previous methods utilizing human feedback. Through comprehensive empirical evaluations across three tasks, we demonstrated the effectiveness of our approach using a diverse range of metrics. Notably, SYNDICOM outperformed ChatGPT on all metrics for both the dialogue improvement tasks, with and without human feedback.

To facilitate further research and practical adoption, we plan to release the code implementation of SYNDICOM as well as the complete dataset utilized in this work. By making these resources openly accessible, we aim to encourage collaboration and promote advancements in commonsense reasoning for dialogue systems.

Limitations and Future Work

There are a few areas of limitation in this work. First, all the dialogues generated were based on templates synthesized from ATOMIC triplets. The domain is thus limited to the material contained in ATOMIC. Second, the procedural generation technique, while scaleable, inevitably introduces structure within the data that can be exploited by statistical models (including deep neural nets and language models). This is why the feedback generation task is particularly crucial, because the explanations are human-written and thus avoid such a limitation.

Our experiments demonstrate our method of improving baseline dialogue responses that have been corrupted with error injection. This has the advantage of scale and targeting specific error modes that may be observed with LLMs, but the invalid responses in SYNDICOM do not themselves represent errors actually made by LLMs. A larger scale study could involve a data collection of errors and mistakes made by an LLM to demonstrate our method in improving baseline dialogue responses, but this approach would not lend itself to scale as any particular type of error made by state-of-the-art LLMs will likely be very rare. A more scaleable approach might be to develop a more comprehensive error typology and injection scheme, which we leave to future work.

In future work, a more comprehensive error topology could be explored, along with a more substantial human evaluation, to explore the generalizability of the proposed method. This work focused on commonsense errors, but other errors that are observed in large language models could be explored in further analysis like mathematical reasoning, humor and sarcasm, etc.

References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022b. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. 2020. Mutual: A dataset for multi-turn dialogue reasoning. *arXiv preprint arXiv:2004.04494*.

Deepanway Ghosal, Siqi Shen, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. 2022. CICERO: A dataset for contextualized commonsense inference in dialogues. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5010–5028, Dublin, Ireland. Association for Computational Linguistics.

Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6384–6392.

Hyunwoo Kim, Youngjae Yu, Liwei Jiang, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi, and Maarten Sap. 2022. Prosocialdialog: A prosocial backbone for conversational agents. *arXiv preprint arXiv:2205.12688*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

Bing Liu, Gokhan Tur, Dilek Hakkani-Tur, Pararth Shah, and Larry Heck. 2017. End-to-end optimization of task-oriented dialogue model with deep reinforcement learning. *Conversational AI Workshop, Neural Information Processing Systems (NeurIPS)*.

Bing Liu, Gokhan Tür, Dilek Hakkani-Tür, Pararth Shah, and Larry Heck. 2018. Dialogue learning with human teaching and feedback in end-to-end trainable task-oriented dialogue systems. 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 2060–2069, New Orleans, Louisiana. Association for Computational Linguistics.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *arXiv preprint arXiv:1506.08909*.

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 845–854.

Neha Nayak, Dilek Hakkani-Tür, Marilyn A Walker, and Larry P Heck. 2017. To plan or not to plan? discourse planning in slot-value informed sequence to sequence models for language generation. In *INTERSPEECH*, pages 3339–3343.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Lianhui Qin, Aditya Gupta, Shyam Upadhyay, Luheng He, Yejin Choi, and Manaal Faruqui. 2021. Timedial: Temporal commonsense reasoning in dialog. *arXiv preprint arXiv:2106.04571*.

Christopher Richardson and Larry Heck. 2023. Commonsense reasoning for conversational ai: A survey of the state of the art. *arXiv preprint arXiv:2302.07926*.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3027–3035.

William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. 2022. Self-critiquing models for assisting human evaluators. *arXiv preprint arXiv:2206.05802*.

Jérémy Scheurer, Jon Ander Campos, Jun Shern Chan, Angelica Chen, Kyunghyun Cho, and Ethan Perez. 2022. Training language models with natural language feedback. *arXiv preprint arXiv:2204.14146*.

Pararth Shah, Dilek Hakkani-Tür, Tür, and Larry Heck. 2016. Interactive reinforcement learning for task-oriented dialogue management. *Workshop on Deep Learning for Action and Interaction, Neural Information Processing Systems (NIPS)*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Hao Sun, Zhexin Zhang, Fei Mi, Yasheng Wang, Wei Liu, Jianwei Cui, Bin Wang, Qun Liu, and Minlie Huang. 2022. Moraldial: A framework to train and evaluate moral dialogue systems via constructing moral discussions. *arXiv preprint arXiv:2212.10720*.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. Dream: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231.

Sean Welleck, Ximing Lu, Peter West, Faeze Brahman, Tianxiao Shen, Daniel Khashabi, and Yejin Choi. 2022. Generating sequences by learning to self-correct. *arXiv preprint arXiv:2211.00053*.

Jason Weston, Emily Dinan, and Alexander H Miller. 2018. Retrieve and refine: Improved sequence generation models for dialogue. *arXiv preprint arXiv:1808.04776*.

Pei Zhou, Karthik Gopalakrishnan, Behnam Hedayatnia, Seokhwan Kim, Jay Pujara, Xiang Ren, Yang Liu, and Dilek Hakkani-Tur. 2021a. Commonsense-focused dialogues for response generation: An empirical study. *arXiv preprint arXiv:2109.06427*.

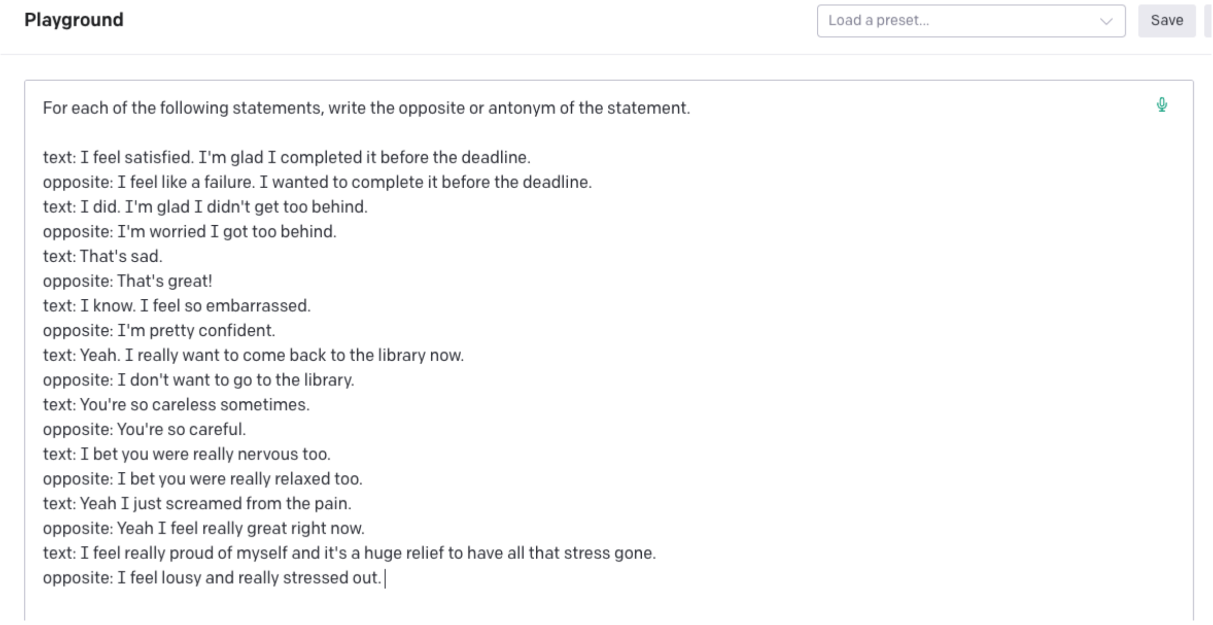
Ruijie Zhou, Soham Deshmukh, Jeremiah Greer, and Charles Lee. 2021b. Narle: Natural language models using reinforcement learning with emotion feedback. *arXiv preprint arXiv:2110.02148*.

Caleb Ziems, Jane A Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2022. The moral integrity corpus: A benchmark for ethical dialogue systems. *arXiv preprint arXiv:2204.03021*.

A GPT-3 Prompts and Mechanical Turk interfaces

Task	Prompt
Direct	You will be given a dialogue context and a baseline response. Your job is to improve that baseline response. Always write the improved response last and prefix it with 'Improved Response:'
NLHF	You will be given a dialogue context and a baseline response. Your job is to improve that baseline response. Do so by first generating feedback for that response, as if it was written by an AI and you are critiquing it, and then produce the improved response. Always write the improved response last and prefix it with 'Improved Response:'
Feedback Generation	You are shown a synthetic dialogue written by an AI. The dialogue is intended to sound like a natural text message conversation between two people. The AI is imperfect and makes mistakes. You are asked to provide feedback to the AI to improve its dialogue generation. You are given a few dialogue turns, followed by a Baseline Response. Please give 1-2 sentences of feedback for the baseline response, and please be specific!

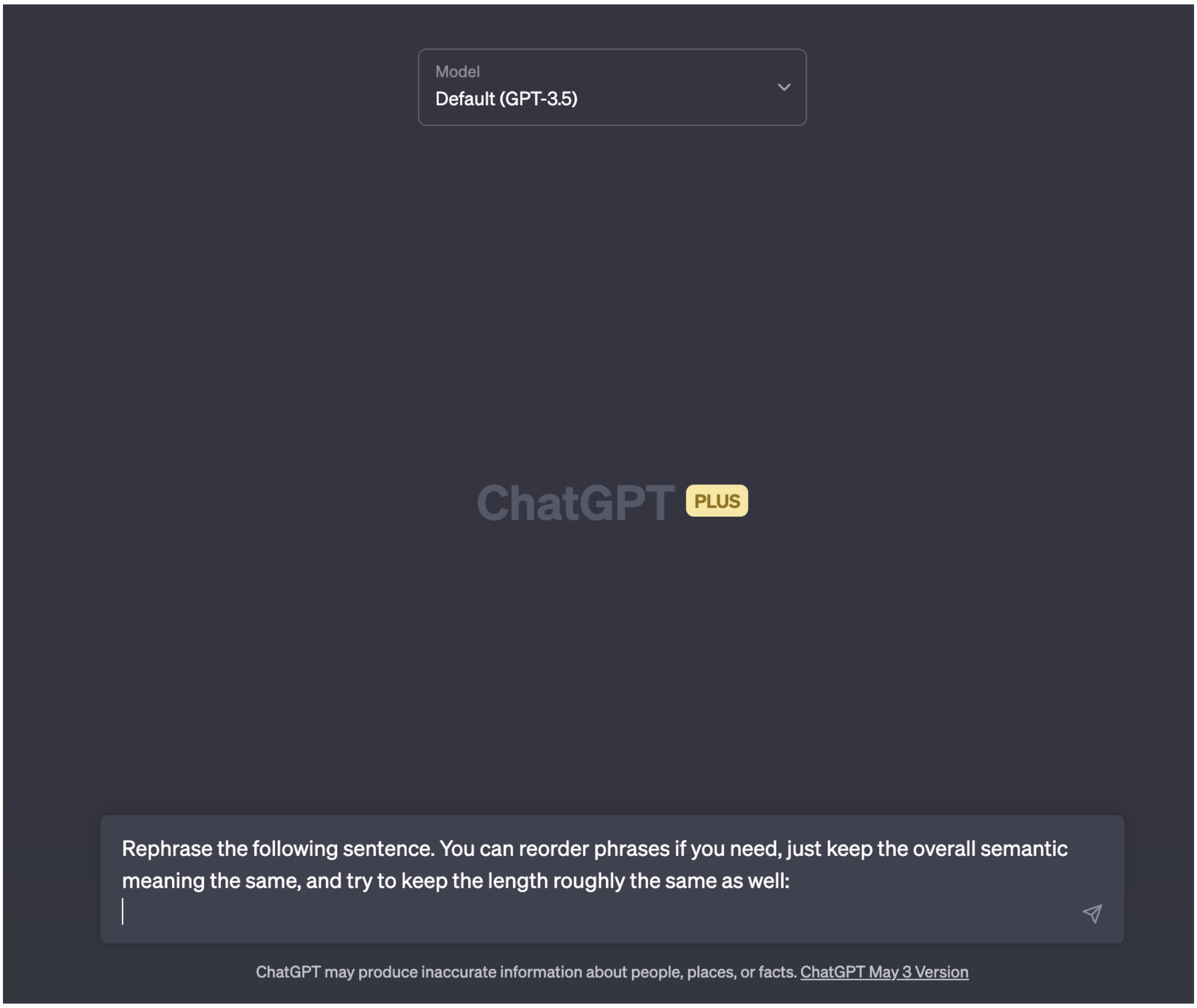
Table 5: Prompts used for ChatGPT baselines



For each of the following statements, write the opposite or antonym of the statement.

text: I feel satisfied. I'm glad I completed it before the deadline.
opposite: I feel like a failure. I wanted to complete it before the deadline.
text: I did. I'm glad I didn't get too behind.
opposite: I'm worried I got too behind.
text: That's sad.
opposite: That's great!
text: I know. I feel so embarrassed.
opposite: I'm pretty confident.
text: Yeah. I really want to come back to the library now.
opposite: I don't want to go to the library.
text: You're so careless sometimes.
opposite: You're so careful.
text: I bet you were really nervous too.
opposite: I bet you were really relaxed too.
text: Yeah I just screamed from the pain.
opposite: Yeah I feel really great right now.
text: I feel really proud of myself and it's a huge relief to have all that stress gone.
opposite: I feel lousy and really stressed out.

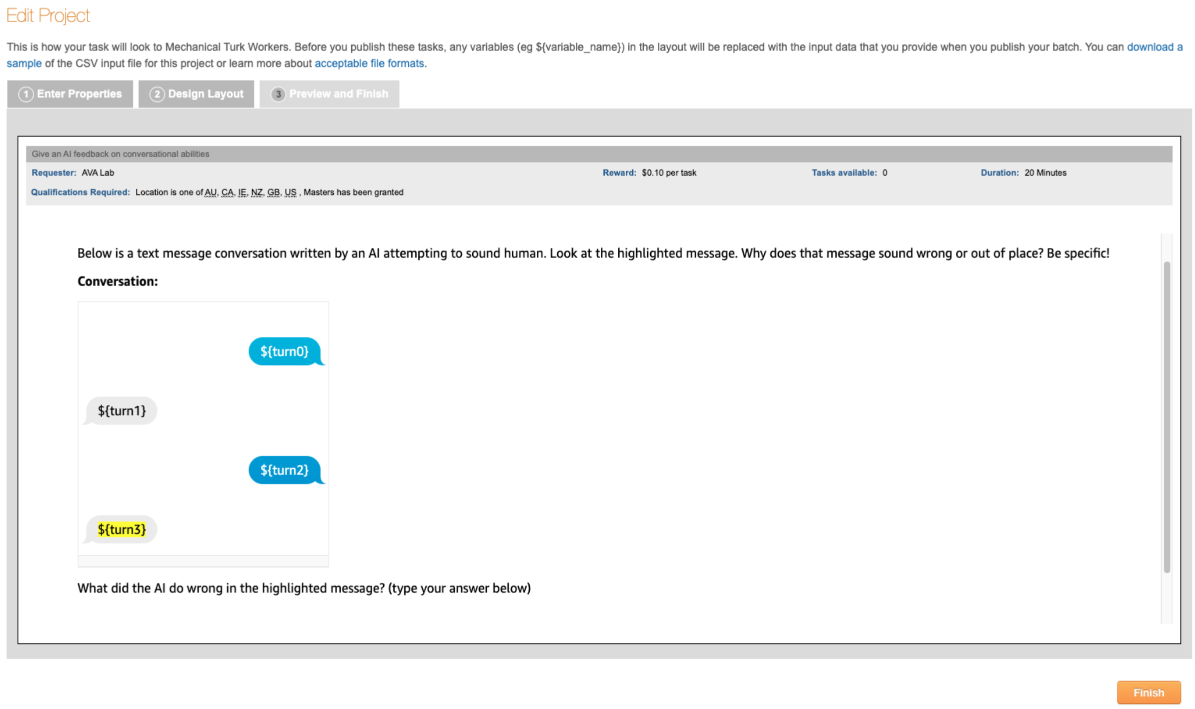
Figure 2: GPT-3 Prompt used for creating invalid dialogue responses from valid responses.



Rephrase the following sentence. You can reorder phrases if you need, just keep the overall semantic meaning the same, and try to keep the length roughly the same as well:

ChatGPT may produce inaccurate information about people, places, or facts. ChatGPT May 3 Version

Figure 3: ChatGPT prompt used for rephrasing invalid dialogue responses.



Edit Project

This is how your task will look to Mechanical Turk Workers. Before you publish these tasks, any variables (eg ${variable_name}) in the layout will be replaced with the input data that you provide when you publish your batch. You can download a sample of the CSV input file for this project or learn more about acceptable file formats.

Give an AI feedback on conversational abilities

Requester: AVA Lab Reward: $0.10 per task Tasks available: 0 Duration: 20 Minutes

Qualifications Required: Location is one of AU, CA, IE, NZ, GB, US , Masters has been granted

Below is a text message conversation written by an AI attempting to sound human. Look at the highlighted message. Why does that message sound wrong or out of place? Be specific!

Conversation:

${turn0}

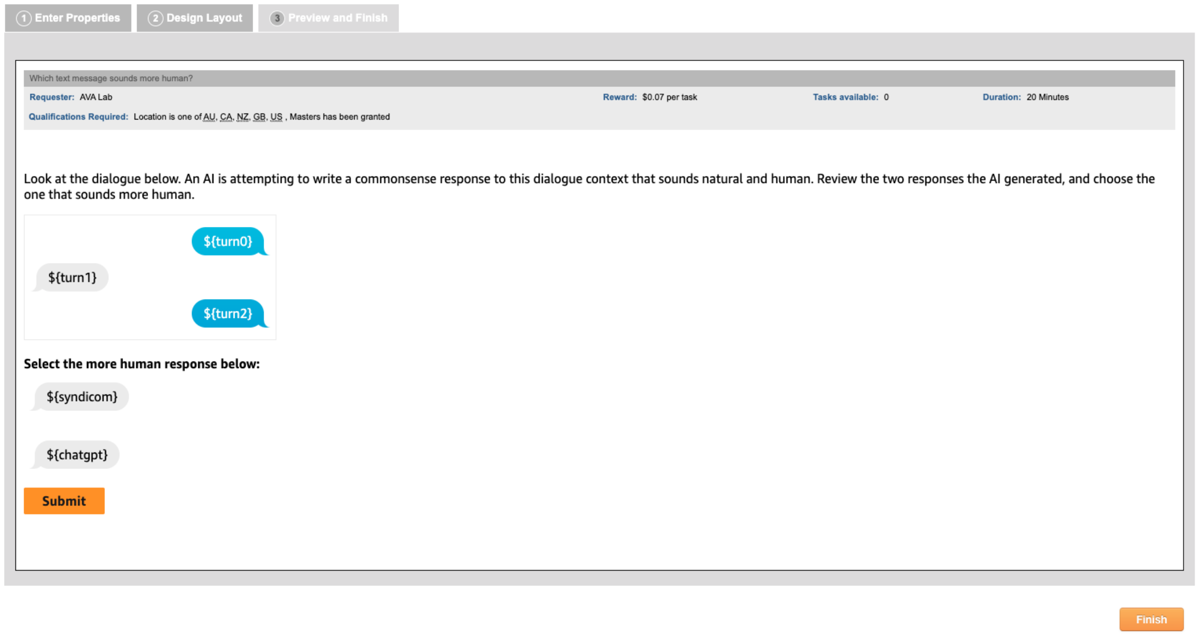
${turn1}

${turn2}

${turn3}

What did the AI do wrong in the highlighted message? (type your answer below)

Figure 4: Mechanical Turk interface used for acquiring feedback for dialogue responses. Each dialogue was given feedback by two independent crowdworkers.



Which text message sounds more human?

Requester: AVA Lab Reward: $0.07 per task Tasks available: 0 Duration: 20 Minutes

Qualifications Required: Location is one of AU, CA, NZ, GB, US , Masters has been granted

Look at the dialogue below. An AI is attempting to write a commonsense response to this dialogue context that sounds natural and human. Review the two responses the AI generated, and choose the one that sounds more human.

${turn0}

${turn1}

${turn2}

Select the more human response below:

${syndicom}

${chatgpt}

Figure 5: Mechanical Turk interface used for human evaluation. Each dialogue response pair was evaluated by two workers independently. Templates are shown instead of examples in order to fit the page.