Ditch the Gold Standard: Re-evaluating Conversational Question Answering

Huihan Li* **Tianyu Gao*** **Manan Goenka** **Danqi Chen**
Department of Computer Science, Princeton University
{huihanl,tianyug,mgoenka,danqic}@princeton.edu

Abstract

Conversational question answering aims to provide natural-language answers to users in information-seeking conversations. Existing conversational QA benchmarks compare models with pre-collected human-human conversations, using ground-truth answers provided in conversational history. It remains unclear whether we can rely on this static evaluation for model development and whether current systems can well generalize to real-world human-machine conversations. In this work, we conduct the first large-scale human evaluation of state-of-the-art conversational QA systems, where human evaluators converse with models and judge the correctness of their answers. We find that the distribution of human-machine conversations differs drastically from that of human-human conversations, and there is a disagreement between human and gold-history evaluation in terms of model ranking. We further investigate how to improve automatic evaluations, and propose a question rewriting mechanism based on predicted history, which better correlates with human judgments. Finally, we analyze the impact of various modeling strategies and discuss future directions towards building better conversational question answering systems.[1]

1 Introduction

Conversational question answering aims to build machines to answer questions in conversations and has the promise to revolutionize the way humans interact with machines for information seeking. With recent development of large-scale datasets (Choi et al., 2018; Saeidi et al., 2018; Reddy et al., 2019; Campos et al., 2020), rapid progress has been made in better modeling of conversational QA systems.

Current conversational QA datasets are collected by crowdsourcing human-human conversations, where the questioner asks questions about a specific topic, and the answerer provides answers based on an evidence passage and the conversational history. When evaluating conversational QA systems, a set of held-out conversations are used for asking models questions in turn. Since the evaluation builds on pre-collected conversations, the *gold history* of the conversation is always provided, regardless of models' actual predictions (Figure 1(b)). Although current systems achieve near-human F1 scores on this static evaluation, it is questionable whether this can faithfully reflect models' true performance in real-world applications. To what extent do human-machine conversations deviate from human-human conversations? What will happen if models have no access to ground-truth answers in a conversation?

To answer these questions and better understand the performance of conversational QA systems, we carry out the first large-scale human evaluation with four state-of-the-art models on the QuAC dataset (Choi et al., 2018) by having human evaluators converse with the models and judge the correctness of their answers. We collected 1,446 human-machine conversations in total, with 15,059 question-answer pairs. Through careful analysis, we notice a significant distribution shift from human-human conversations and identify a clear inconsistency of model performance between current evaluation protocol and human judgements.

This finding motivates us to improve automatic evaluation such that it is better aligned with human evaluation. Mandya et al. (2020); Siblini et al. (2021) identify a similar issue in gold-history evaluation and propose to use models' own predictions for automatic evaluation. However, predicted-history evaluation poses another challenge: since all the questions have been collected beforehand, using predicted history will invalidate some of the questions because of changes in the conversational history (see Figure 1(c) for an example).

Following this intuition, we propose a *question*

*The first two authors contributed equally.
[1]Our data and code are publicly available at https://github.com/princeton-nlp/EvalConvQA.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 8074 - 8085

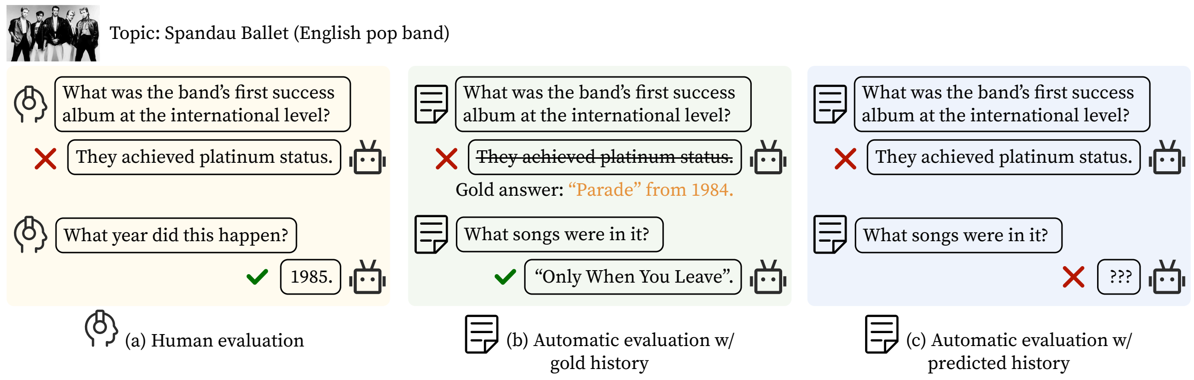


Figure 1: Examples of *human* and *automatic* evaluations with gold or predicted history. The model answers the first question incorrectly. (a) A human questioner asks the next question based on current predictions. (b) Automatic evaluation with gold history poses pre-collected questions with gold answers as conversational history, regardless of model predictions. (c) Using predicted history in automatic evaluation makes the next question invalid.

rewriting mechanism, which automatically detects and rewrites invalid questions with predicted history (Figure 4). We use a coreference resolution model (Lee et al., 2018) to detect inconsistency of conference in question text conditioned on predicted history and gold history, and then rewrite those questions by substituting with correct mentions, so that the questions are resolvable in the predicted context. Compared to predicted-history evaluation, we find that incorporating this rewriting mechanism aligns better with human evaluation.

Finally, we also investigate the impact of different modeling strategies based on human evaluation. We find that both accurately detecting unanswerable questions and explicitly modeling question dependencies in conversations are crucial for model performance. Equipped with all the insights, we discuss directions for conversational QA modeling. We release our human evaluation dataset and hope that our findings can shed light on future development of better conversational QA systems.

2 Preliminary

2.1 Evaluation of conversational QA

Evaluation of conversational QA in real-world consists of three components: an evidence passage P, a (human) questioner $\mathcal{H}$ that has no access to P,[2] and a model $\mathcal{M}$ that has access to P. The questioner asks questions about P and the model answers them based on P and the conversational history thus far (see an example in Figure 1(a)). Formally, for the i-th turn, the human asks a question based on the previous conversation,

$$Q_i \sim \mathcal{H}(Q_1, A_1, ..., Q_{i-1}, A_{i-1}), \tag{1}$$

and then the model answers it based on both the history and the passage,

$$A_i \sim \mathcal{M}(P, Q_1, A_1, ..., Q_{i-1}, A_{i-1}, Q_i), \tag{2}$$

where Q_i and A_i represent the question and the answer at the i-th turn. If the question is unanswerable from P, we simply denote A_i as `CANNOT ANSWER`. The model $\mathcal{M}$ is then evaluated by the correctness of answers.

Evaluating conversational QA systems requires human in the loop and is hence expensive. Instead, current benchmarks use automatic evaluation with gold history (*Auto-Gold*) and collect a set of human-human conversations for automatic evaluation. For each passage, one annotator asks questions without seeing the passage, while the other annotator provides the answers. Denote the collected questions and answers as Q_i^* and A_i^*. In gold-history evaluation, the model is inquired with pre-collected questions Q_i^* and the gold answers as history:

$$A_i \sim \mathcal{M}(P, Q_1^*, A_1^*, ..., Q_{i-1}^*, A_{i-1}^*, Q_i^*), \tag{3}$$

and we evaluate the model by comparing A_i to A_i^* (measured by word-level F1). This process does not require human effort but cannot truly reflect the distribution of human-machine conversations, because unlike human questioners who may ask different questions based on different model predictions, this static process ignores model predictions and always asks the pre-collected question.

In this work, we choose the QuAC dataset (Choi et al., 2018) as our primary evaluation because it is

[2] Existing conversational QA datasets make different assumptions: For example, QuAC (Choi et al., 2018) assumes no access but CoQA assumes the questioner to have access.

closer to real-world information-seeking conversations, where the questioner *cannot* see the evidence passage during the dataset collection. It prevents the questioner asking questions that simply overlaps with the passage and encourages unanswerable questions. QuAC also adopts *extractive* question answering that restricts the answer as a span of text, which is generally considered easier to evaluate.

2.2 Models

For human evaluation and analysis, we choose the following four conversational QA models with different model architectures and training strategies:

BERT. It is a simple BERT (Devlin et al., 2019) baseline which concatenates the previous two turns of question-answer pairs, the question, and the passage as the input and predicts the answer span.[3] This model is the same as the "BERT + PHQA" baseline in Qu et al. (2019a).

GraphFlow. Chen et al. (2020) propose a recurrent graph neural network on top of BERT embeddings to model the dependencies between the question, the history and the passage.

HAM. Qu et al. (2019b) propose a history attention mechanism (HAM) to softly select the most relevant previous turns.

ExCorD. Kim et al. (2021) train a question rewriting model on CANARD (Elgohary et al., 2019) to generate context-independent questions, and then use both the original and the generated questions to train the QA model. This model achieves the current state-of-the-art on QuAC (67.7% F1).

For all the models except BERT, we use the original implementations for a direct comparison. We report their performance on both standard benchmark and our evaluation in Table 2.

3 Human Evaluation

3.1 Conversation collection

In this section, we carry out a large-scale human evaluation with the four models discussed above. We collect human-machine conversations using 100 passages from the QuAC development set on Amazon Mechanical Turk.[4] We also design a set of qualification questions to make sure that the annotators fully understand our annotation guideline. For each model and each passage, we collect three conversations from three different annotators.

We collect each conversation in two steps:

(1) The annotator has no access to the passage and asks questions. The model extracts the answer span from the passage or returns `CANNOT ANSWER` in a human-machine conversation interface.[5] We provide the title, the section title, the background of the passage, and the first question from QuAC as a prompt to annotators. Annotators are required to ask at least 8 and at most 12 questions. We encourage context-dependent questions, but also allow open questions like "What else is interesting?" if asking a follow-up question is difficult. (2) After the conversation ends, the annotator is shown the passage and asked to check whether the model predictions are correct or not.

We noticed that the annotators are biased when evaluating the correctness of answers. For questions to which the model answered `CANNOT ANSWER`, annotators tend to mark the answer as incorrect without checking if the question is answerable. Additionally, for answers with the correct types (e.g. a date as an answer to "When was it?"), annotators tend to mark it as correct without verifying it from the passage. Therefore, we asked another group of annotators to verify question answerability and answer correctness.

3.2 Answer validation

For each collected conversation, we ask two additional annotators to validate the annotations. First, each annotator reads the passage before seeing the conversation. Then, the annotator sees the question (and question only) and selects whether the question is (a) ungrammatical, (b) unanswerable, or (c) answerable. If the annotator chooses "answerable", the interface then reveals the answer and asks about its correctness. If the answer is "incorrect", the annotator selects the correct answer span from the passage. We discard all questions that both annotators find "ungrammatical" and the correctness is taken as the majority of the 3 annotations.

3.3 Annotation statistics

In total, we collected 1,446 human-machine conversations and 15,059 question-answer pairs. We release this collection as an important source that

[3] We use `bert-base-uncased` as the encoder.

[4] We restrict the annotators from English-speaking countries, and those who have finished at least 1,000 HITS with an acceptance rate of >95%. The compensation rate for Amazon Mechanical Turk workers is calculated using $15/h.

[5] We used ParlAI (Miller et al., 2017) to build the interface.

	Human Evaluation				QuAC
	BERT	GF	HAM	ExCorD	
# C	357	359	373	357	1,000
# Q	3,755	3,666	3,959	3,679	7,354

Table 1: Number of conversations (# C) and questions (# Q) collected in human evaluation, using 100 passages from the QuAC development set. We also add QuAC *development* set for reference. GF: GraphFlow.

complements existing conversational QA datasets. Numbers of conversations and question-answer pairs collected for each model are shown in Table 1. The data distribution of this collection is very different from the original QuAC dataset (human-human conversations): we see more open questions and unanswerable questions, due to less fluent conversation flow caused by model mistakes, and that models cannot provide feedback to questioner about whether an answer is worth following up like human answerers do (more analysis in §6.2).

Deciding the correctness of answers is challenging even for humans in some cases, especially when questions are short and ambiguous. We measure annotators' agreement and calculate the Fleiss' Kappa (Fleiss, 1971) on the agreement between annotators in the validation phase. We achieve $\kappa = 0.598$ (moderate agreement) of overall annotation agreement. Focusing on answerability annotation, we have $\kappa = 0.679$ (substantial agreement).

4 Disagreements between Human and Gold-history Evaluation

We now compare the results from our human evaluation and gold-history (automatic) evaluation. Note that the two sets of numbers are not directly comparable: (1) the human evaluation reports accuracy, while the automatic evaluation reports F1 scores; (2) the absolute numbers of human evaluation are much higher than those of automatic evaluations. For example, for the BERT model, the human evaluation accuracy is 82.6% while the automatic evaluation F1 is only 63.2%. The reason is that, in automatic evaluations, the gold answers cannot capture all possible correct answers to open-ended questions or questions with multiple answers; however, the human annotators can evaluate the correctness of answers easily. Nevertheless, we can compare relative rankings between different models.

Figure 2 shows different trends between human evaluation and gold-history evaluation (Auto-Gold).

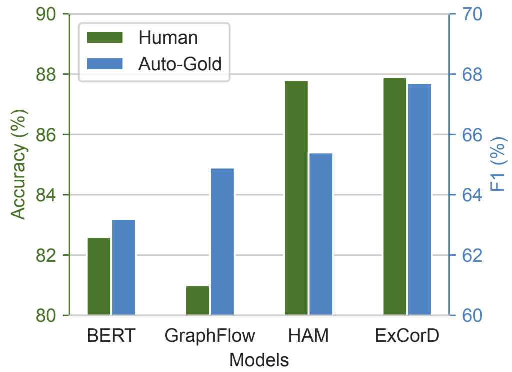


Figure 2: Model performance of human evaluation (accuracy, left) and Auto-Gold (F1, right). Scales for accuracy and F1 are different. Human evaluation and Auto-Gold rank BERT and GraphFlow differently.

Current standard evaluation cannot reflect model performance in human-machine conversations: (1) Human evaluation and Auto-Gold rank BERT and GraphFlow differently; especially, GraphFlow performs much better in automatic evaluation, but worse in human evaluation. (2) The gap between HAM and ExCorD is significant (F1 of 65.4% vs 67.7%) in the automatic evaluation but the two models perform similarly in human evaluation (accuracy of 87.8% vs 87.9%).

5 Strategies for Automatic Evaluation

The inconsistency between human evaluation and gold-history evaluation suggests that we need better ways to evaluate and develop our conversational QA models. When being deployed in realistic scenarios, the models would never have access to the ground truth (gold answers) in previous turns and are only exposed to the conversational history and the passage. Intuitively, we can simply replace gold answers by the predicted answers of models and we name this as **predicted-history** evaluation (*Auto-Pred*). Formally, the model makes predictions based on the questions and its own answers:

$$A_i \sim \mathcal{M}(P, Q_1^*, A_1, ..., Q_{i-1}^*, A_{i-1}, Q_i^*). \quad (4)$$

This evaluation has been suggested by several recent works (Mandya et al., 2020; Siblini et al., 2021), which reported a significant performance drop using predicted history. We observe the same performance degradation, shown in Table 2.

However, another issue naturally arises with predicted history: Q_i^*s were written by the dataset annotators based on $(Q_1^*, A_1^*, ..., Q_{i-1}^*, A_{i-1}^*)$, which

Unresolved coreference (44.0%)	
Q_1^*:	What was Frenzal Rhomb's first song?
A_1^*:	Punch in the Face.
A_1:	CANNOT ANSWER.
Q_2^*:	How did *it* fare?
Incoherence (39.1%)	
Q_1^*:	Did Billy Graham succeed in becoming a chaplain?
A_1^*:	He *contracted mumps* shortly after...
A_1:	After a period of recuperation in Florida, he ...
Q_2^*:	Did he retire after his *mumps diagnosis*?
Correct answer changed (16.9%)	
Q_1^*:	Are there any other interesting aspects?
A_1^*:	... *Steve Di Giorgio returned to the band*...
A_1:	... bassist Greg Christian had left Testament again...
Q_2^*:	What happened following *this change in crew*?

Figure 3: Examples of invalid questions with predicted history. Some are shortened for better demonstration. Q_i^*, A_i^*: questions and gold answers from the collected dataset, A_i: model predictions.

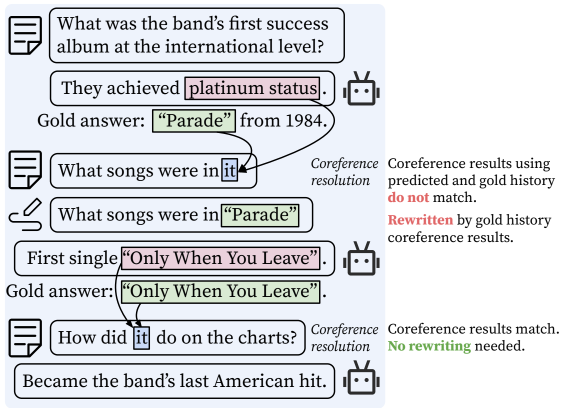


Figure 4: An example of question rewriting. We rewrite the second question with referent in the gold history, because predicted and gold history have different coreference results. We do not rewrite the third question as coreference results are the same.

may become unnatural or invalid when the history is changed to $(Q_1^*, A_1, ..., Q_{i-1}^*, A_{i-1})$.

5.1 Predicted history invalidates questions

We examined 100 QuAC conversations with the best-performing model (ExCorD) and identified three categories of invalid questions caused by predicted history. We find that 23% of the questions become invalid after using the predicted history. We summarize the types of invalid questions as follows (see detailed examples in Figure 3):

- **Unresolved coreference** (44.0%). The question becomes invalid for containing either a pronoun or a definite noun phrase that refers to an entity unresolvable without the gold history.
- **Incoherence** (39.1%). The question is incoherent with the conversation flow (e.g., mentioning an entity non-existent in predicted history). While humans may still answer the question using the passage, this leads to an unnatural conversation and a train-test discrepancy for models.
- **Correct answer changed** (16.9%). The answer to this question with the predicted history changes from when it is based on the gold history.

We further analyze the reasons for the biggest "unresolved coreference" category and find that the model either gives an incorrect answer to the previous question ("incorrect prediction", 39.8%), or the model predicts a different (yet correct) answer to an open question ("open question", 37.0%), or the model returns CANNOT ANSWER incorrectly ("no prediction", 9.5%), or the gold answer is longer than prediction and the next question depends on the extra part ("extra gold information", 13.6%). Invalid questions result in compounding errors, which may further affect how the model interprets the following questions.

5.2 Evaluation with question substitution

Among all the invalid question categories, "unresolved coreference" questions are the most critical ones. They lead to incorrect interpretations of questions and hence wrong answers. We propose to improve our evaluation by first detecting these questions using a state-of-the-art coreference resolution system (Lee et al., 2018)[6], and then substituting them with either rewriting the questions in-place and replacing the questions with their context-independent counterparts.

Detecting invalid questions. We make the assumption that if the coreference model resolves mentions in Q_i^* differently between using gold history $(Q_1^*, A_1^*, ..., A_{i-1}^*, Q_i^*)$ and predicted history $(Q_1^*, A_1, ..., A_{i-1}, Q_i^*)$, then Q_i^* is identified as having an unresolved coreference issue.

The inputs to the coreference model for Q_i^* are the following:

$$S_i^* = [BG; Q_{i-k}^*; A_{i-k}^*; Q_{i-k+1}^*; A_{i-k+1}^*; ...; Q_i^*]$$
$$S_i = [BG; Q_{i-k}^*; A_{i-k}; Q_{i-k+1}^*; A_{i-k+1}; ...; Q_i^*],$$

[6] We use the coreference model from AllenNLP (Gardner et al., 2018).

	All				Answerable questions			
	BERT	GraphFlow	HAM	ExCorD	BERT	GraphFlow	HAM	ExCorD
Auto-Gold (F1)	63.2	64.9	65.4	67.7	61.8	66.6	64.5	66.4
Auto-Pred (F1)	54.6	49.6	57.2	61.2	52.7	54.5	54.6	59.2
Auto-Rewrite (F1)	54.5	48.2	57.3	61.9	51.2	51.9	55.1	59.7
Auto-Replace (F1)	54.2	47.8	57.1	61.7	50.7	51.7	54.8	59.7
Human (Accuracy)	82.6	81.0	87.8	87.9	75.9	83.2	84.8	85.3

Table 2: Model performance in automatic and human evaluations. We report *overall performance* on all questions and also performance on *answerable questions* only.

where BG is the background, S_i^* and S_i denote the inputs for gold and predicted history. After the coreference model returns entity cluster information given S_i^* and S_i, we extract a list of entities $E^* = \{e_1^*, ..., e_{|E^*|}^*\}$ and $E = \{e_1, ..., e_{|E|}\}$.[7] We say Q_i^* is *valid* only if $E^* = E$, that is,

$$|E^*| = |E| \text{ and } e_j^* = e_j, \forall e_j \in E,$$

assuming e_j^* and e_j have a shared mention in Q_i^*. We determine whether $e_j^* = e_j$ by checking if $\mathrm{F1}(s_j^*, s_j) > 0$, where s_j^* and s_j are the *first* mention of e_j^* and e_j respectively, and F1 is the word-level F1 score, i.e., $e_j^* = e_j$ as long as their first mentions have word overlap. The reason we take the F1 instead of exact match to check whether the entities are the same is stated in Appendix A.

Question rewriting through entity substitution. Our first strategy is to substitute the entity names in Q_i^* with entities in E^*, if Q_i^* is invalid. The rewritten question, instead of the original one, will be used in the conversation history and fed into the model. We denote this evaluation method as **rewritten-question** evaluation (*Auto-Rewrite*), and Figure 4 illustrates a concrete example.

To analyze how well Auto-Rewrite does in detecting and rewriting questions, we manually check 100 conversations of ExCorD from the QuAC development set. We find that Auto-Rewrite can detect invalid questions with a precision of 72% and a recall of 72% (more detailed analysis in Appendix B). An example of correctly detected and rewritten question is presented in Figure 4.

Question replacement using CANARD. Another strategy is to replace the invalid questions with context-independent questions. The CANARD dataset (Elgohary et al., 2019) provides such a resource, which contains human-rewritten context-independent version of QuAC's questions. Recent works (Anantha et al., 2021; Elgohary et al., 2019) have proposed training sequence-to-sequence models on such dataset to rewrite questions; however, since the performance of the question-rewriting models is upper bounded by the human-rewritten version, we simply use CANARD for question replacement. We denote this strategy as **replaced-question** evaluation (*Auto-Replace*). Because collecting context-independent questions is expensive, Auto-Replace is limited to evaluating models on QuAC; it is also possible to be extended to other datasets by training a question rewriting model, as demonstrated in existing work.

[7]We are only interested in the entities mentioned in the current question Q_i^* and we filter out named entities (e.g., the *National Football League*) because they can be understood without coreference resolution.

6 Automatic vs Human Evaluation

In this section, we compare human evaluation with all the automatic evaluations we have introduced: gold-history (Auto-Gold), predicted-history (Auto-Pred), and our proposed Auto-Rewrite and Auto-Replace evaluations. We first explain the metrics we use in the comparison (§6.1) and then discuss the findings (§6.2 and §6.3).

6.1 Agreement metrics

Model performance and rankings. We first consider using model performance reported by different evaluation methods. Considering numbers of automatic and human evaluations are not directly comparable, we also calculate models' rankings and compare whether the rankings are consistent between automatic and human evaluations. Model performance is reported in Table 2. In human evaluation, GraphFlow $<$ BERT $<$ HAM $\approx$ ExCorD; in Auto-Gold, BERT $<$ GraphFlow $<$ HAM $<$ ExCorD; in other automatic evaluations, GraphFlow $<$ BERT $<$ HAM $<$ ExCorD.

Human Evaluation				QuAC
BERT	GF	HAM	ExCorD	
34.6	20.6	34.1	33.2	20.2

Table 3: Percentage of unanswerable questions in human evaluation (it varies with different models) and the original QuAC dataset (used in all automatic evaluations). GF: GraphFlow.

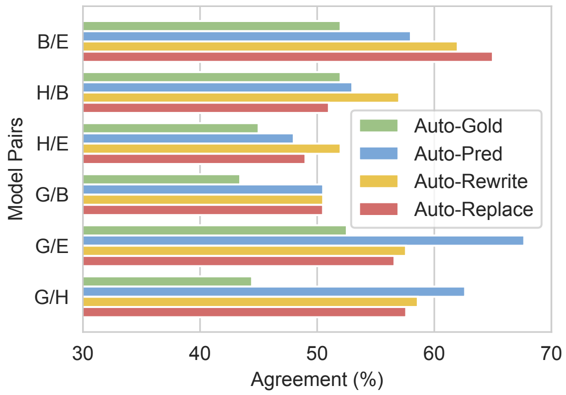


Figure 5: Pairwise agreement of different model pairs comparing automatic evaluations to human evaluation. B: BERT; G: GraphFlow; H: HAM; E: ExCorD.

Statistics of unanswerable questions. Percentage of unanswerable questions is an important aspect in conversations. Automatic evaluations using static datasets have a fixed number of unanswerable questions, while in human evaluation, the percentage of unanswerable questions asked by human annotators varies with different models. The statistics of unanswerable questions is shown in Table 3.

Pairwise agreement. For a more fine-grained evaluation, we perform a passage-level comparison for every pair of models. More specifically, for every single passage we use one automatic metric to decide whether model A outperforms model B (or vice versa) and examine the percentage of passages that the automatic metric agrees with human evaluation. For example, if the pairwise agreement of BERT/ExCorD between human evaluation and Auto-Gold is 52%, it means that Auto-Gold and human evaluation agree on 52% passages in terms of which model is better. Higher agreement means the automatic evaluation is closer to human evaluation. Figure 5 shows the results of pairwise agreement.

6.2 Automatic evaluations have a significant distribution shift from human evaluation

We found that automatic evaluations have a significant distribution shift from human evaluation. We draw this conclusion from the following points.

- Human evaluation shows a much higher model performance than all automatic evaluations, as shown in Table 2. Two reasons may cause this large discrepancy: (a) Many conversational QA questions have multiple possible answers, and it is hard for the static dataset in automatic evaluations to capture all the answers. It is not an issue in human evaluation because all answers are judged by human evaluators. (b) There are more unanswerable questions and open questions in human evaluation (reason discussed in the next paragraph), which are easier—for example, models are almost always correct when answering questions like "What else is interesting?".

- Human evaluation has a much higher unanswerable question rate, as shown in Table 3. The reason is that in human-human data collection, the answers are usually correct and the questioners can ask followup questions upon the high-quality conversation; in human-machine interactions, since the models can make mistakes, the conversation flow is less fluent and it is harder to have followup questions. Thus, questioners chatting with models tend to ask more open or unanswerable questions.

- All automatic evaluation methods have a pairwise agreement lower than 70% with human evaluation, as shown in Figure 2. This suggests that all automatic evaluations cannot faithfully reflect the model performance of human evaluation.

6.3 Auto-Rewrite is closer to human evaluation

First, we can clearly see that among all automatic evaluations, Auto-Gold deviates the most from the human evaluation. From Table 2, only Auto-Gold shows different rankings from human evaluation, while Auto-Pred, Auto-Rewrite, and Auto-Replace show consistent rankings to human judgments.

In Figure 2, we see that Auto-Gold has the lowest agreement with human evaluation; among others, Auto-Rewrite better agrees with human evaluation for most model pairs. Surprisingly, Auto-Rewrite is even better than Auto-Replace—which uses human-written context independent questions—in most cases. After checking the Auto-Replace conversations, we found that human-written context independent questions are usually much longer than QuAC questions and introduce extra information

	Predicted unanswerable Q.				Precision				Recall			
	B	G	H	E	B	G	H	E	B	G	H	E
Auto-Gold	27.1	21.5	27.1	28.3	56.8	62.3	57.1	57.9	68.1	59.3	68.4	72.5
Auto-Pred	27.8	13.8	28.6	28.9	50.0	53.9	52.3	53.3	61.4	33.0	66.1	68.2
Auto-Rewrite	27.3	13.1	25.1	26.0	48.6	55.0	52.4	53.9	65.7	35.7	65.1	69.4
Auto-Replace	27.5	12.9	25.2	25.7	48.6	54.2	52.1	53.8	66.1	34.7	64.9	68.4
Human	42.3	14.7	37.2	36.0	75.0	93.0	86.8	87.4	95.2	72.5	93.7	93.3

Table 4: The percentage of models' predicted unanswerable questions, and the precision and recall for detecting unanswerable questions in different evaluations. B: BERT; G: GraphFlow; H: HAM; E: ExCorD.

into the context, which leads to out-of-domain challenges for conversational QA models (example in Appendix C). It shows that our rewriting strategy can better reflect real-world performance of conversational QA systems. However, Auto-Rewrite is not perfect—we see that when comparing G/E or G/H, Auto-Pred is better than Auto-Rewrite; in all model pairs, the agreement between human evaluation and Auto-Rewrite is still lower than 70%. This calls for further effort in designing better automatic evaluation in the future.

7 Towards Better Conversational QA

With insights drawn from human evaluation and comparison with automatic evaluations, we discuss the impact of different modeling strategies, as well as future directions towards building better conversational question answering systems.

Modeling question dependencies on conversational context. When we focus on *answerable questions* (Table 2), we notice that GraphFlow, HAM and ExCorD perform much better than BERT. We compare the modeling differences of the four systems in Figure 6, and identify that all the three better systems explicitly model the question dependencies on the conversation history and the passage: both GraphFlow and HAM highlight repeated mentions in questions and conversation history by special embeddings (turn marker and PosHAE) and use attention mechanism to select the most relevant part from the context; ExCorD adopts a question rewriting module that generates context-independent questions given the history and passage. All those designs help models better understand the question in a conversational context. Figure 7 gives an example where GraphFlow, HAM and ExCorD resolved the question from long conversation history while BERT failed.

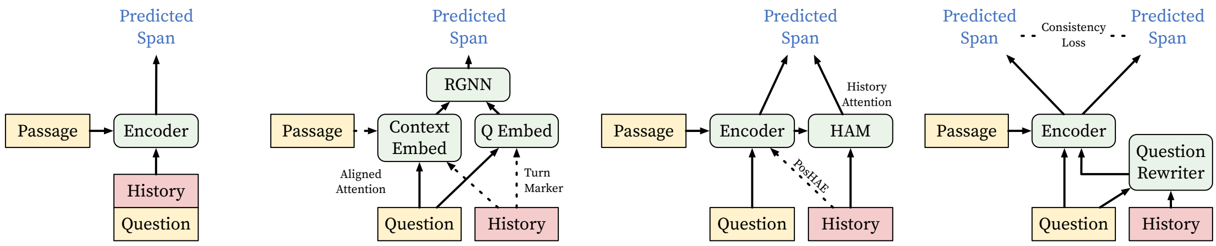
Unanswerable question detection. Table 4 demonstrates models' performance in detecting *unanswerable questions*. We notice that GraphFlow predicts much fewer unanswerable questions than the other three models, and has a high precision and a low recall in unanswerable detection. This is because GraphFlow uses a separate network for predicting unanswerable questions, which is harder to calibrate, while the other models jointly predict unanswerable questions and answer spans.

This behavior has two effects: (a) GraphFlow's overall performance is dragged down by its poor unanswerable detection result (Table 2). (b) In human evaluation, annotators ask fewer unanswerable questions with GraphFlow (Table 3)—when the model outputs more, regardless of correctness, the human questioner has a higher chance to ask passage-related followup questions. Both suggest that how well the model detects unanswerable questions significantly affects its performance and the flow in human-machine conversations.

Optimizing towards the new testing protocols. Most existing works on conversational QA modeling focus on optimizing towards Auto-Gold evaluation. Since Auto-Gold has a large gap from the real-world evaluation, more efforts are needed in optimizing towards the human evaluation, or Auto-Rewrite, which better reflects human evaluation. One potential direction is to improve models' robustness given noisy conversation history, which simulates the inaccurate history in real world that consists of models' own predictions. In fact, prior works (Mandya et al., 2020; Siblini et al., 2021) that used predicted history in training showed that it benefits the models in predicted-history evaluation.

8 Related Work

Conversational question answering. In recent years, several conversational question answering datasets have emerged, such as QuAC (Choi



(a) BERT (b) GraphFlow (c) HAM (d) ExCorD

Figure 6: Modeling structures of BERT, GraphFlow, HAM, and ExCorD.

Tom McCall – Vortex I
... McCall decided to hold a rock festival at Milo McIver State Park, Oregon called "Vortex I: A Biodegradable Festival of Life"...

Q_1^*:	Was Vortex I popular?	
B:	*The festival*, "The Governor's Pot Party" ...	✓
G/H/E:	*The festival*, "The Governor's Pot Party" ...	✓
...		
Q_4^*:	Who played at *the festival*?	
B:	`CANNOT ANSWER`	✗
G/H/E:	Gold, The Portland Zoo, Osceola, Fox...	✓

Figure 7: An example of BERT failing to resolve *the festival* in Q_4^*, while all other models with explicit dependency modelings succeeded.

et al., 2018), CoQA (Reddy et al., 2019), and DoQA (Campos et al., 2020), as well as a few recent works focusing on conversational open-domain question answering (Adlakha et al., 2021; Anantha et al., 2021; Qu et al., 2020) Different from single-turn QA datasets (Rajpurkar et al., 2016), conversational QA requires the model to understand the question in the context of conversational history. There have been many methods proposed to improve conversational QA performance (Ohsugi et al., 2019; Chen et al., 2020; Qu et al., 2019b; Kim et al., 2021) and significant improvements have been made on conversational QA benchmarks. Besides text-based conversational QA tasks, there also exist conversational QA benchmarks that require external knowledge or other modalities (Saeidi et al., 2018; Saha et al., 2018; Guo et al., 2018; Das et al., 2017).

Only recently has it been noticed that the current method of evaluating conversational QA models is flawed. Mandya et al. (2020); Siblini et al. (2021) point out that using gold answers in history is not consistent with real-world scenarios and propose to use predicted history for evaluation. Different from prior works, in this paper, we conduct a large scale human evaluation to provide evidence for why gold-history evaluation is sub-optimal. In addition, we point out that even predicted-history evaluation has issues with invalid questions, for which we propose rewriting questions to further mitigate the gap.

Automatic evaluation of dialogue systems. Automatically evaluating dialogue systems is difficult due to the nature of conversations. In recent years, the NLP community has cautiously re-evaluated and identified flaws in many popular automated evaluation strategies of dialogue systems (Liu et al., 2016; Sai et al., 2019), and have proposed new evaluation protocols to align more with human evaluation in a real-world setting: Huang et al. (2020); Ye et al. (2021) evaluate the coherence of the dialogue systems; Gupta et al. (2019) explore to use multiple references for evaluation; Mehri and Eskenazi (2020) propose an unsupervised and reference-free evaluation; Lowe et al. (2017); Tao et al. (2018); Ghazarian et al. (2019); Shimanaka et al. (2019); Sai et al. (2020) train models to predict the relatedness score between references and model outputs, which are shown to be better than BLEU (Papineni et al., 2002) or ROGUE (Lin, 2004).

9 Conclusion

In this work, we carry out the first large-scale human evaluation on conversational QA systems. We show that current standard automatic evaluation with gold history cannot reflect models' performance in human evaluation, and that human-machine conversations have a large distribution shift from static conversational QA datasets of human-human conversations. To tackle these problems, we propose to use predicted history with rewriting invalid questions for evaluation, which reduces the gap between automatic evaluations and real-world human evaluation. Based on the insights from the human evaluation results, we also nalyze current conversational QA systems and identify promising directions for future development.

Acknowledgements

We thank Alexander Wettig and other members of the Princeton NLP group, and the anonymous reviewers for their valuable feedback. This research is supported by a Graduate Fellowship at Princeton University and the James Mi *91 Research Innovation Fund for Data Science.

References

Vaibhav Adlakha, Shehzaad Dhuliawala, Kaheer Suleman, Harm de Vries, and Siva Reddy. 2021. Topiocqa: Open-domain conversational question answering with topic switching. *arXiv preprint arXiv:2110.00768*.

Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2021. Open-domain question answering goes conversational via question rewriting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 520–534, Online. Association for Computational Linguistics.

Jon Ander Campos, Arantxa Otegi, Aitor Soroa, Jan Deriu, Mark Cieliebak, and Eneko Agirre. 2020. DoQA - accessing domain-specific FAQs via conversational QA. In *Association for Computational Linguistics (ACL)*, pages 7302–7314.

Yu Chen, Lingfei Wu, and Mohammed J Zaki. 2020. Graphflow: Exploiting conversation flow with graph neural networks for conversational machine comprehension. In *International Joint Conference on Artificial Intelligence (IJCAI)*.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 2174–2184.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1080–1089.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186.

Ahmed Elgohary, Denis Peskov, and Jordan Boyd-Graber. 2019. Can you unpack that? learning to rewrite questions-in-context. In *Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5918–5924.

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

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6.

Sarik Ghazarian, Johnny Tian-Zheng Wei, Aram Galstyan, and Nanyun Peng. 2019. Better automatic evaluation of open-domain dialogue systems with contextualized embeddings. *arXiv preprint arXiv:1904.10635*.

Daya Guo, Duyu Tang, Nan Duan, Ming Zhou, and Jian Yin. 2018. Dialog-to-action: Conversational question answering over a large-scale knowledge base. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2942–2951.

Prakhar Gupta, Shikib Mehri, Tiancheng Zhao, Amy Pavel, Maxine Eskenazi, and Jeffrey Bigham. 2019. Investigating evaluation of open-domain dialogue systems with human generated multiple references. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 379–391, Stockholm, Sweden. Association for Computational Linguistics.

Lishan Huang, Zheng Ye, Jinghui Qin, Liang Lin, and Xiaodan Liang. 2020. GRADE: Automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems. In *emnlp*, pages 9230–9240.

Gangwoo Kim, Hyunjae Kim, Jungsoo Park, and Jaewoo Kang. 2021. Learn to resolve conversational dependency: A consistency training framework for conversational question answering. In *Association for Computational Linguistics (ACL)*, pages 6130–6141.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 687–692.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*.

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic Turing test: Learning to evaluate dialogue responses. In *Association for Computational Linguistics (ACL)*, pages 1116–1126. Association for Computational Linguistics.

Angrosh Mandya, James O' Neill, Danushka Bollegala, and Frans Coenen. 2020. Do not let the history haunt you: Mitigating compounding errors in conversational question answering. In *International Conference on Language Resources and Evaluation (LREC)*, pages 2017–2025.

Shikib Mehri and Maxine Eskenazi. 2020. USR: An unsupervised and reference free evaluation metric for dialog generation. In *Association for Computational Linguistics (ACL)*, pages 681–707.

Alexander Miller, Will Feng, Dhruv Batra, Antoine Bordes, Adam Fisch, Jiasen Lu, Devi Parikh, and Jason Weston. 2017. ParlAI: A dialog research software platform. In *Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations*, pages 79–84.

Yasuhito Ohsugi, Itsumi Saito, Kyosuke Nishida, Hisako Asano, and Junji Tomita. 2019. A simple but effective method to incorporate multi-turn context with BERT for conversational machine comprehension. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 11–17.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Association for Computational Linguistics (ACL)*, pages 311–318.

Chen Qu, Liu Yang, Cen Chen, Minghui Qiu, W. Bruce Croft, and Mohit Iyyer. 2020. *Open-Retrieval Conversational Question Answering*, page 539–548. Association for Computing Machinery, New York, NY, USA.

Chen Qu, Liu Yang, Minghui Qiu, W Bruce Croft, Yongfeng Zhang, and Mohit Iyyer. 2019a. Bert with history answer embedding for conversational question answering. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 1133–1136.

Chen Qu, Liu Yang, Minghui Qiu, Yongfeng Zhang, Cen Chen, W Bruce Croft, and Mohit Iyyer. 2019b. Attentive history selection for conversational question answering. In *ACM International Conference on Information and Knowledge Management (CIKM)*, pages 1391–1400.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 2383–2392.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association of Computational Linguistics (TACL)*, pages 249–266.

Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. 2018. Interpretation of natural language rules in conversational machine reading. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 2087–2097.

Amrita Saha, Vardaan Pahuja, Mitesh M. Khapra, Karthik Sankaranarayanan, and Sarath Chandar. 2018. Complex sequential question answering: Towards learning to converse over linked question answer pairs with a knowledge graph. *arXiv preprint arXiv:1801.10314*.

Ananya B Sai, Mithun Das Gupta, Mitesh M Khapra, and Mukundhan Srinivasan. 2019. Re-evaluating adem: A deeper look at scoring dialogue responses. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6220–6227.

Ananya B. Sai, Akash Kumar Mohankumar, Siddhartha Arora, and Mitesh M. Khapra. 2020. Improving dialog evaluation with a multi-reference adversarial dataset and large scale pretraining. *Transactions of the Association of Computational Linguistics (TACL)*, 8:810–827.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2019. Machine translation evaluation with bert regressor. *arXiv preprint arXiv:1907.12679*.

Wissam Siblini, Baris Sayil, and Yacine Kessaci. 2021. Towards a more robust evaluation for conversational question answering. In *Association for Computational Linguistics and International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 1028–1034.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Zheng Ye, Liucun Lu, Lishan Huang, Liang Lin, and Xiaodan Liang. 2021. Towards quantifiable dialogue coherence evaluation. In *acl*, pages 2718–2729. Association for Computational Linguistics.

A Invalid Question Detection

In question rewriting, we use F1 instead of exact match to check whether two entites are the same. The reason is that sometimes the prediction may mention the same entity as the gold answer does, but with different names. Figure 8 gives an example. Thus to avoid the false positive of detecting invalid questions, we take the F1 metric.

Q_1^*: Who is at the door?
A_1^*: *An elderly Chinese lady* and a little boy
A_1: *elderly Chinese lady*

Q_2^*: Is *she* carrying something?

Figure 8: An example that the prediction may mention the same entity as the gold answer does with slightly different names.

B Quality of Rewriting Questions

Detection. After manually checking 100 conversations of ExCorD from the QuAC development set, we find that Auto-Rewrite can detect invalid questions with a precision of 72% and a recall of 72%. We notice that the coreference model sometimes detects the pronoun of the main character in the passage as insolvable, although it almost shows up in every question. This issue causes the low precision but is not a serious problem in our case – whether rewriting the pronoun of the main character does not affect models' prediction much, because the model always sees the passage and knows who the main character is.

Rewriting. Among all correctly detected invalid questions, we further check the quality of rewriting, and in 68% of the times Auto-Rewrite gives a correct context-independent questions. The most common error is being ungrammatical. For example, using the gold history of "... Dee Dee claimed that Spector once ***pulled*** a gun on him", the original question "Did they arrest him for doing *this*?" was rewritten to "Did they arrest Phillip Harvey Spector for doing ***pulled***?" While this creates a distribution shift on question formats, it is still better than putting an invalid question in the flow.

C Issue with Context Independent Questions

Figure 9 shows an example where extra information in context-independent questions confuses the model and leads to incorrect prediction.

Q^*:	Did he go on to any other notable matches?
Q^W:	Did *he* go on to any other notable matches?
A^W:	During the Test match series against Australia in 2010, at the Melbourne Cricket Ground...
Q^P:	Did *Mohammad Amir* go on to any other notable matches, *besides on 9 November 2009*?
A^P:	Later in 2009, Pakistan toured Sri Lanka

Figure 9: The context-independent question Q^P by Auto-Replace contains extra information that confuses the model. The rewritten question Q^W did not change the original question and led to a correct answer.