Realistic Conversational Question Answering with Answer Selection based on Calibrated Confidence and Uncertainty Measurement

Soyeong Jeong[1] **Jinheon Baek**[2] **Sung Ju Hwang**[1,2] **Jong C. Park**[1*]
School of Computing[1] Graduate School of AI[2]
Korea Advanced Institute of Science and Technology[1,2]
{starsuzi,jinheon.baek,sjhwang82,jongpark}@kaist.ac.kr

Abstract

Conversational Question Answering (ConvQA) models aim at answering a question with its relevant paragraph and previous question-answer pairs that occurred during conversation multiple times. To apply such models to a real-world scenario, some existing work uses predicted answers, instead of unavailable ground-truth answers, as the conversation history for inference. However, since these models usually predict wrong answers, using all the predictions without filtering significantly hampers the model performance. To address this problem, we propose to filter out inaccurate answers in the conversation history based on their estimated confidences and uncertainties from the ConvQA model, without making any architectural changes. Moreover, to make the confidence and uncertainty values more reliable, we propose to further calibrate them, thereby smoothing the model predictions. We validate our models, Answer Selection-based realistic Conversation Question Answering, on two standard ConvQA datasets, and the results show that our models significantly outperform relevant baselines. Code is available at: https://github.com/starsuzi/AS-ConvQA.

1 Introduction

Conversational Question Answering (ConvQA) is the task of answering a series of questions during conversation, taking into account a given relevant paragraph (Choi et al., 2018; Reddy et al., 2019). Contrary to traditional extractive question answering tasks (Rajpurkar et al., 2016; Trischler et al., 2017) that answer each question with the given paragraph just once, ConvQA aims at answering the current question using its previous question-answer pairs taking into account the given paragraph multiple times. For example, as illustrated in Figure 1, the goal of ConvQA is to correctly answer the question Q_3 based on the previous conversation history such as Q_2, A_2, Q_1, and A_1, as well as the current context C.

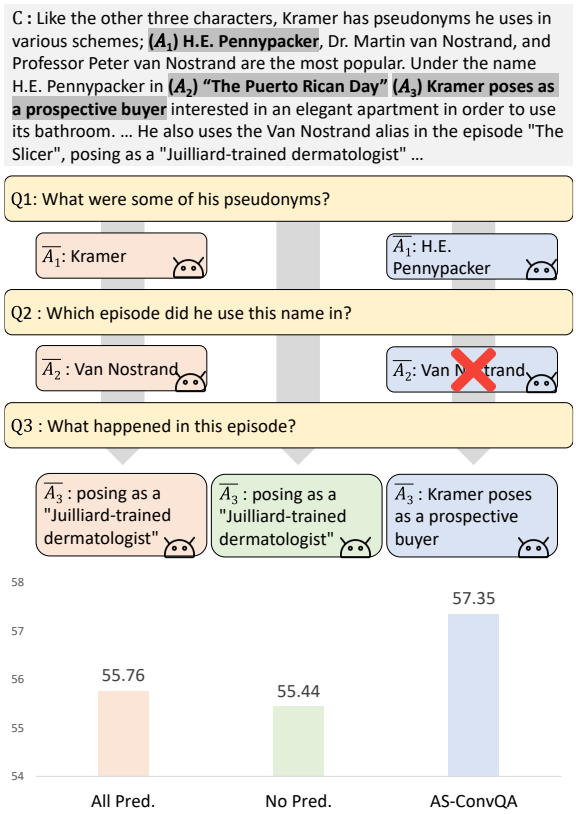

Figure 1: Illustration of realistic ConvQA evaluation with three models: 1) using all predicted answers (All Pred.); 2) not using predicted answers (No Pred.); 3) only using probably correct answers while filtering out others (AS-ConvQA, Ours). The scores in the bar chart underneath represent the F1 scores measured by all test samples (see Table 1 for full results).

ConvQA has recently gained much attention as it follows the human's information seeking process through multi-turn interactions with others. However, it is also known to be quite challenging since it requires capturing all the information over the current question, previous conversation, and the given paragraph. To tackle this problem, a considerable amount of work focuses mainly on developing a model architecture for ConvQA (Qu et al., 2019a,b; Huang et al., 2019; Chen et al., 2020; Kim et al., 2021; Qiu et al., 2021; Raposo et al., 2022).

Despite their successes, however, there remains

* Corresponding author

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 477–490
May 2-6, 2023

a critical limitation in that they use the ground-truth answers (i.e., A_2 and A_1) in the conversation history during both training and evaluation steps. Such an evaluation procedure is not applicable to the real-world scenario, since the ground-truth answers are not accessible when the user's query is posed. Therefore, the supporting dialogue history for the current question should consist of the model's predictions $\bar{A}_1$ and $\bar{A}_2$ in the real-world application, instead of the nonexistent gold answers A_1 and A_2.

There is some recent work (Mandya et al., 2020; Siblini et al., 2021) that considers such a realistic setting on evaluation. In particular, they propose to use the model's predicted answers (i.e., $\bar{A}_1$ and $\bar{A}_2$), instead of the ground-truth answers (i.e., A_1 and A_2), for its evaluation. However, in such a setting, the model faces inconsistency between training and evaluation since the model is evaluated with the predictions while trained with the ground-truth answers. To handle such a discrepancy, Mandya et al. (2020) and Siblini et al. (2021) suggest strategies that randomly decide whether to use predicted or gold answers for the input question during training.

However, as Figure 1 shows, using all predictions as the answer history is not effective: The performance difference is not so significant when compared to not using them at all. We see that this originates from a model's failure to answering previous questions. Specifically, if a model incorrectly predicts an answer $\bar{A}_1$ for the previous question Q_1, using the incorrectly predicted answer $\bar{A}_1$ for the question Q_2 not only affects the model's current prediction $\bar{A}_2$ negatively, but also engenders further errors in the future prediction for Q_3.

Therefore, in this work, we propose a novel selection scheme for predicted answers from the conversation history, which filters out predictions that are likely to be incorrect, unlike the existing work that uses all the predicted answers including incorrect ones. The remaining step is then to identify possibly incorrect predictions. To this end, we propose to use the confidence and uncertainty of the model's prediction, which are measured by its likelihood and entropy, respectively. In particular, if the model predicts the previous answer with lower confidence (i.e., lower likelihood) or higher uncertainty (i.e., higher entropy) than a certain threshold, we regard the model's previous answer as probably incorrect, and remove it from the conversation history in answering the current question during evaluation. On the other hand, during training, we soften the sampling process so that, instead of using the hard threshold above, we sample a predicted answer based on its confidence or uncertainty (e.g., the lower the uncertainty, the higher the chance to include the predicted answer in the conversation history), in order to diversify the model's input.

However, when dealing with confidence and uncertainty, we should be careful about a miscalibrated situation (Guo et al., 2017), which happens when uncertainty and confidence do not correspond to the error and accuracy of ground-truth correctness, respectively. In other words, if the model is not calibrated enough and the distribution for confidence and uncertainty is highly skewed over particular ranges, the highly uncertain or low confident yet valid predictions could be removed. Therefore, to prevent such a performance degrading situation, we further calibrate models using a temperature scaling scheme (Guo et al., 2017) before estimating the uncertainty or confidence. We refer to our method as Answer Selection-based realistic Conversational Question Answering (AS-ConvQA).

We validate our method on two standard ConvQA datasets, QuAC (Choi et al., 2018) and CoQA (Reddy et al., 2019), against diverse baselines on a realistic evaluation protocol. The experimental results show that our method significantly outperforms these baselines, and a detailed analysis supports the importance of uncertainty- and confidence-based answer selection schemes.

Our contributions in this work are threefold:

- We propose to remove incorrect predictions in a conversation history, which degenerate ConvQA models' performances during inference.
- We present confidence- and uncertainty-based answer filtering schemes, which are further calibrated to obtain reliable predictions.
- We show that our method achieves outstanding performances on realistic ConvQA tasks.

2 Related Work

Conversational Question Answering ConvQA requires a model to understand the context of questions and paragraphs along with previous conversational questions and answers (Choi et al., 2018; Reddy et al., 2019). While the simplest approach to consider such conversation histories is to embed them along with the given question and paragraph in the representation space, recent work (Huang et al., 2019; Qu et al., 2019b; Chen et al., 2020)

proposed to leverage the relevant histories by selectively using them. However, Vakulenko et al. (2021) and Kim et al. (2021) have shown that even a simple concatenation of previous questions and answers outperforms these selection-based methods, thanks to the advances in pre-trained language models (Devlin et al., 2019; Liu et al., 2019) that are designed to attend to the relevant parts. Furthermore, recent methods (Elgohary et al., 2019; Kim et al., 2021; Vakulenko et al., 2021; Raposo et al., 2022) rather focus on the problem of ambiguity in the input question by proposing a question rewriting scheme for its disambiguation, showing remarkable performance improvements.

However, the aforementioned work has a fundamental limitation on the evaluation protocol: They evaluate models based on ground-truth answers working as a conversation history, which are not available in a real-world setting. Li et al. (2022) point out this problem of ground-truth history evaluation but use the ground-truth answers during evaluation as well, since they target at disambiguating pronouns in the question by comparing the predicted and ground-truth answers. Alternatively, Mandya et al. (2020) and Siblini et al. (2021) use the model's predictions instead of the ground-truth answers during evaluation, and further train the model with predicted answers. However, they do not take into account the quality of predicted answers, where low-quality ones are not useful (see Figure 1). Thus, we propose to selectively use the predicted answers that are probably correct, based on their calibrated confidences and uncertainties.

Confidence and Uncertainty As it is nearly impossible for models to always make accurate predictions, unreliable predictions become serious issues when deploying machine learning models to real-world settings. Motivated to prevent such a risk, mechanisms of estimating the reliability of model's predictive probabilities based on confidence and uncertainty are recently proposed (Abdar et al., 2021; Houben et al., 2022). We note that confidence is usually measured by the softmax outputs of models (Guo et al., 2017), and that uncertainty can be quantified by Bayesian models, which can be approximated via Monte Carlo (MC) dropout (Gal and Ghahramani, 2016; Kendall and Gal, 2017). With much work on confidence and uncertainty estimations in computer vision tasks (Guillory et al., 2021), related topics have been recently adopted for NLP tasks as well (Shelmanov et al., 2021; Wu et al., 2021; Malinin and Gales, 2021; Vazhentsev et al., 2022). While confidence and uncertainty estimation should also be considered in ConvQA, we believe that this venue is under-explored so far. In particular, since questions are asked sequentially, it is likely that untrustworthy predictions in the conversation history would negatively affect the performance. To tackle this, we propose to exclude low-confident or uncertain predictions when training and evaluating the ConvQA model.

Calibration Confidence and uncertainty help interpret the validity of the model's prediction. However, it is not safe to rely on them when the model is not calibrated, where the correct likelihood does not match the predicted probability (Guo et al., 2017), or the model error does not match the predicted uncertainty (Laves et al., 2019). Since deep neural networks are prone to miscalibration as the number of parameters has much increased, large pre-trained language models are also not free from this problem (Wang et al., 2021; Zhao et al., 2021; Dan and Roth, 2021). One of the most prevalent approaches to calibrating the model is to rescale a logit vector before the softmax function for regularizing the probability, which is known as temperature scaling (Guo et al., 2017). While there exist lots of calibration schemes, including label smoothing (Szegedy et al., 2016) and confidence penalty (Pereyra et al., 2017), in this work, we use temperature scaling as a calibrator, since it is simple yet effective while not changing the output class of the model prediction (i.e., only scaling logits).

3 Method

We first introduce ConvQA. Then, we describe our answer validating methods based on confidence and uncertainty values with their calibration schemes.

3.1 Conversational Question Answering

We provide general descriptions of a ConvQA task. For the i-th turn of the conversation, we are given a question Q_i and its corresponding context C, as well as its conversation history consisting of previous questions and answers: $\mathcal{H}_i = \{Q_{i-1}, A_{i-1}, ..., Q_1, A_1\}$. Then, the goal of ConvQA is to correctly extract the ground-truth answer A_i from C along with Q_i and $\mathcal{H}_i$, as follows:

$$P(A_i) = M_\theta(C, Q_i, Q_{i-1}, A_{i-1}, ..., Q_1, A_1), \quad (1)$$

where M_θ is a ConvQA model, parameterized by θ, and, for the sake of simplicity, we omit conditional

variables C, Q_i, and $\mathcal{H}_i$ on the left side of Equation 1, i.e., $P(A_i) = P(A_i|C, Q_i, \mathcal{H}_i)$. Note that existing work (Elgohary et al., 2019; Kim et al., 2021; Vakulenko et al., 2021; Raposo et al., 2022) has an unrealistic assumption that a set of ground-truth answers $\{A_{i-1}, ..., A_1\}$ is available during evaluation as in Equation 1. However, this evaluation setup is far from reality, since they are not always available when the user's novel questions come in, unlike the training phase which optimizes a model with ground-truth answers. Therefore, we should particularly modify the formulation in Equation 1 to accommodate a realistic evaluation scenario, which we describe in the next subsection.

3.2 Realistic ConvQA

To tackle the problem of accessing ground-truth answers during evaluation in Equation 1, we aim at redefining its formulation to evaluate ConvQA models under real-world situations as shown below.

Evaluation When a user asks a unique question whose ground-truth answers are not accessible, the most naïve approach is to work with the relevant context and previous questions, as follows:

$$P(\bar{A}_i) = M_\theta(C, Q_i, Q_{i-1}, ..., Q_1), \quad (2)$$

where $\bar{A}_i$ denotes the i^{th} predicted answer $(i > 1)$ during inference time. However, the formulation in Equation 2 may be suboptimal, since it ignores predicted answers $\{\bar{A}_{i-1}, ..., \bar{A}_1\}$ that occurred in the former conversation, which may be beneficial for the current prediction. Thus, we can instead make inference with predicted answers, as follows:

$$P(\bar{A}_i) = M_\theta(C, Q_i, Q_{i-1}, \bar{A}_{i-1}, ..., Q_1, \bar{A}_1). \quad (3)$$

However, when evaluating with Equation 3 while training with Equation 1, a problematic discrepancy arises, as model M_θ uses gold answers A_i for training but predicted answers $\bar{A}_i$ for inference.

Training To tackle this inconsistency, recent work (Mandya et al., 2020; Siblini et al., 2021) randomly decides whether to use $\bar{A}_i$ or A_i during the training phase, as follows:

$$P(A_i) = \begin{cases} M_\theta(C, Q_i, Q_{i-1}, \bar{A}_{i-1}, ...) \text{ w.p. } \lambda_{rand}, \\ M_\theta(C, Q_i, Q_{i-1}, A_{i-1}, ...) \text{ w.p. } 1 - \lambda_{rand}, \end{cases} \quad (4)$$

where λ_{rand} is the probability of using $\bar{A}_i$, which is set based on heuristic sampling schemes, either using the random coin flipping or increasing the sampling rate based on the number of steps.

While such an attempt bridges the gap between training and inference in the real-world setting, critical limitations remain. First, as Figure 1 shows, we observe that using all the predicted answers rarely contributes to the model performance, as they include incorrect answers that hinder accurate predictions for the current question. Also, we further point out that there still exists a discrepancy between training and evaluation: The model observes ground-truth answers in Equation 4 which are yet unobservable for evaluation in Equation 3. Therefore, to tackle these challenges, we propose to selectively use the predicted answers based on the predictions' confidences and uncertainties.

3.3 Predicted Answer Selection Scheme

Our key intuition is that confidence and uncertainty are simple yet effective measures to filter out inaccurate predictions. Before going into details, we first define the notations. Let $\boldsymbol{x_i} \in X$ be an i^{th} input (i.e., turn) for ConvQA model M_θ, which consists of current question Q_i, its relevant context C, and conversation history $\mathcal{H}_i$. Then, labels of given input $\boldsymbol{x_i}$ are defined as $y_i^{(start)} \in C$ and $y_i^{(end)} \in C$ with $C \in \{1, ..., K\}$, where K is the number of sequence lengths for context C. In other words, $y_i^{(start)}$ and $y_i^{(end)}$ denote the start and end spans, respectively. Further, to predict labels $y_i^{(start)}$ and $y_i^{(end)}$, we first obtain a logit vector $\boldsymbol{z}_i$ for each label[1], and use it for calculating a probability vector $\boldsymbol{p}_i$ over K spans: $\boldsymbol{p}_i = \texttt{softmax}(\boldsymbol{z}_i)$, where $\texttt{softmax}$ is a softmax function.

Confidence We now define the confidence. From the probability $\boldsymbol{p} = \texttt{softmax}(\boldsymbol{z})$, a model likelihood can be interpreted as confidence, as follows[2]:

$$s_{conf} = \max_{y \in C} p(y|\boldsymbol{z}), \quad (5)$$

where s_{conf} denotes the confidence value.

Uncertainty While confidence can estimate how confident the model is on its prediction, it might be also beneficial to measure the model's certainty with Bayesian deep learning techniques (Kendall and Gal, 2017) to prevent erroneous predictions, which we describe here. At first, to calculate the uncertainty value, we need to obtain N different predictions for approximating the model's distribution. To do so, we first enable dropout (Srivastava

[1]We omit superscripts $start$ and end for simplicity.

[2]For simplicity, we omit a turn index i, which is represented in a subscript, for example, Q_i for the i^{th} conversation.

et al., 2014) in the language model during inference, and then forward input $\boldsymbol{x}$ for N times with N different dropout masks, which is referred to as Monte Carlo (MC) dropout (Gal and Ghahramani, 2016). Then, we can obtain probability vector $\boldsymbol{p}$ via MC integration: $\boldsymbol{p} = \frac{1}{N}\sum_{n=1}^{N} \texttt{softmax}(\boldsymbol{z}^{(n)})$, where $\boldsymbol{z}^{(n)}$ is the logit vector from each forward pass. Then, based on probability $\boldsymbol{p}$, the uncertainty is quantified via its entropy over K classes (Kendall and Gal, 2017; Laves et al., 2019), as follows:

$$s_{uncer} = -\frac{1}{\log K}\sum_{k=1}^{K} p(k)\log p(k), \quad (6)$$

where s_{uncer} denotes the uncertainty value, which we normalize to be on a scale between 0 and 1 with $\frac{1}{\log K}$ in Equation 6, following (Laves et al., 2019).

3.4 Calibrating Confidence and Uncertainty

We then describe the calibration schemes to match the model's predicted confidence and uncertainty to its correct likelihood and error, respectively.

Perfect Calibration In order to calibrate trustworthiness of the confidence and uncertainty, we first describe perfectly calibrated situations. Given the input $\boldsymbol{x}$, the model predicts the most likely class, $\bar{y} = \arg\max \boldsymbol{p}$, from the entire classes with the highest probability, $\bar{p} = \max \boldsymbol{p}$. Each perfect calibration for confidence and uncertainty is then as follows (Guo et al., 2017; Laves et al., 2019):

$$\begin{aligned} \mathbb{P}(\bar{y} = y | s_{conf} = p) &= p, \\ \mathbb{P}(\bar{y} \neq y | s_{uncer} = p) &= p, \end{aligned} \quad (7)$$

where y denotes the true label with $\forall p \in [0, 1]$.

Calibration and Uncertainty Error However, perfect calibration defined in Equation 7 is hardly achievable in practical settings due to noise and prediction errors. Thus, we rather define a calibration error to estimate how much the model's prediction is calibrated. One of the most prevalent methods to quantify calibration error for confidence is to measure the difference in expectation between confidence and accuracy as follows (Guo et al., 2017):

$$\mathbb{E}_{s_{conf}}[\,|\,\mathbb{P}(\bar{y} = y | s_{conf} = p) - p|\,], \quad (8)$$

where $\forall p \in [0, 1]$. Also, miscalibration of uncertainty is quantified as follows (Laves et al., 2019):

$$\mathbb{E}_{s_{uncer}}[\,|\,\mathbb{P}(\bar{y} \neq y | s_{uncer} = p) - p|\,]. \quad (9)$$

However, since s_{conf} and s_{uncer} lie in a continuous domain, it is impossible to sample them infinite times for every p when measuring calibration errors. Therefore, we further approximate them in a discrete space, which was in the continuous domain (Equations 8, 9), by dividing the predictions into M bins and then measuring accuracy for each corresponding bin. Formally, accuracy and confidence per bin are as follows (Guo et al., 2017):

$$\begin{aligned} \texttt{acc}(B_m) &= \frac{1}{|B_m|}\sum_{i \in B_m} \mathbf{1}(s_{conf} = y^{(i)}), \\ \texttt{conf}(B_m) &= \frac{1}{|B_m|}\sum_{i \in B_m} s_{conf}, \end{aligned} \quad (10)$$

where B_m is a set of label indices whose values are within the m^{th} bin among M non-overlapping bins. Similarly, error and uncertainty per bin are formally defined as follows:

$$\begin{aligned} \texttt{err}(B_m) &= \frac{1}{|B_m|}\sum_{i \in B_m} \mathbf{1}(s_{uncer} \neq y^{(i)}), \\ \texttt{uncer}(B_m) &= \frac{1}{|B_m|}\sum_{i \in B_m} s_{uncer}. \end{aligned} \quad (11)$$

Using definitions in Equations 10, 11 above, we now measure the approximated calibration errors. Regarding confidence, the Expected Calibration Error (ECE) (Guo et al., 2017) is defined as follows:

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} |\texttt{acc}(B_m) - \texttt{conf}(B_m)|, \quad (12)$$

where n is the number of samples in total. For uncertainty, Expected Uncertainty Calibration Error (UCE) (Laves et al., 2019) is defined as follows:

$$UCE = \sum_{m=1}^{M} \frac{|B_m|}{n} |\texttt{err}(B_m) - \texttt{uncer}(B_m)|. \quad (13)$$

Calibration with Temperature Scaling With the calibration criteria (i.e., ECE and UCE), we now aim at obtaining well-calibrated confidence and uncertainty values having low ECE and UCE. To do so, we apply a temperature scaling scheme, which regulates the scale of the obtained logit vector $\boldsymbol{z}$ with a single scalar, namely temperature $\tau > 0$. Note that temperature scaling does not affect the maximum value of the softmax output; therefore, accuracy is preserved. Formally, the calibrated probability vector $\hat{\boldsymbol{p}}$ is defined as follows:

$$\hat{\boldsymbol{p}} = \texttt{softmax}(\boldsymbol{z}/\tau). \quad (14)$$

We find the τ value based on the low calibration errors, i.e., ECE and UCE, in experiments.

3.5 Overall Pipeline

We now summarize the overall pipeline of our AS-ConvQA framework, which leverages the calibrated confidence and uncertainty values to sample valid predictions for inference, while using them during training as well. Our training pipeline consists of two steps, which we explain below.

Step 1 We start training a model with gold answers A_i following the training protocol in Equation 1, since, if the model cannot observe gold answers, it might fail to capture and generate accurate answers, easily leading to degenerated performances (Mandya et al., 2020). Then, to prepare for Step 2, we make inference with it to obtain prediction $\bar{A}_i$ together with its confidence and uncertainty, for each input $\boldsymbol{x}_i$ in the training set.

Step 2 With the predicted answers and their confidences and uncertainties from Step 1, we further train the model to reflect the predicted answers instead of the ground-truth answers. Note that our objective is to filter out less confident or uncertain predictions in inference. Thus, since filtered ones are not observable during our realistic evaluation phase, we also aim at reflecting such an occurrence during training to narrow the gap between training and evaluation. To do so, instead of training with all predicted answers, we rather sample a predicted answer based on its confidence or uncertainty value:

$$P(A_i) = \begin{cases} M_\theta(C, Q_i, Q_{i-1}, \bar{A}_{i-1}, ...) \text{ w.p. } \lambda_{valid}, \\ M_\theta(C, Q_i, Q_{i-1}, ...) \text{ w.p. } 1 - \lambda_{valid}, \end{cases} \tag{15}$$

where λ_{valid} is obtained by the previous prediction's ($\bar{A}_{i-1}$) confidence or uncertainty: $\lambda_{valid} \in [s_{conf}, 1 - s_{uncer}]$. Note that, in contrast to existing work (Mandya et al., 2020; Siblini et al., 2021) represented in Equation 4, our work does not use previous gold-answers (A_{i-1}) for training as well.

For evaluation, we follow the realistic evaluation protocol described in Equation 3. However, instead of using all predictions (Mandya et al., 2020; Siblini et al., 2021), we rather remove low-confident or uncertain predictions against the threshold.

4 Experimental Setups

We explain datasets, metric, and models. Please see Appendix A for further implementation details.

4.1 Dataset and Metric

QuAC QuAC (Choi et al., 2018) is the benchmark ConvQA dataset, which is known to resemble a realistic information seeking dialogue, where questioners were prevented from reading paragraphs for its collection. QuAC consists of 14K dialogues and 100K pairs of questions and paragraphs. As the test set is not publicly open, we use a development set.

	QuAC		CoQA	
	BERT	**RoBERTa**	**BERT**	**RoBERTa**
Gold	59.86	65.08	72.79	77.62
No Pred.	55.44	61.24	70.83	75.56
All Pred.	55.76	61.53	71.28	75.42
CoQAM	55.83	61.55	71.27	74.29
Robust-P	54.21	60.32	70.17	73.96
Attentive Selection	55.74	61.42	71.05	74.60
AS-ConvQA$_{\texttt{conf}}$ (Ours)	**57.03**	**62.47**	**72.00**	**76.52**
AS-ConvQA$_{\texttt{uncer}}$ (Ours)	**57.35**	**62.33**	**72.08**	**76.33**
AS-ConvQA$_{\texttt{combine}}$ (Ours)	**57.06**	**62.18**	**71.99**	**76.76**

Table 1: F1-scores on QuAC and CoQA. Note that Gold model is not a fair baseline as it uses the ground-truth answers during inference, and thus is evaluated in an unrealistic setting.

CoQA CoQA (Reddy et al., 2019) is another ConvQA dataset with 127K pairs of questions and paragraphs; however, unlike QuAC, questioners were allowed to share paragraphs during collection. We also use a development set instead of the test set, which is not publicly available.

F1-score We evaluate models with F1-score, following the standard protocol (Kim et al., 2021).

4.2 Question Answering Models

For question answering models, we use two base-size pre-trained language models widely used in ConvQA tasks: **BERT-base** (Devlin et al., 2019) and **RoBERTa-base** (Liu et al., 2019).

4.3 Baselines and Our Models

We compare AS-ConvQA to other relevant baselines using predicted answers. Gold model, which is an indicator, uses gold answers as the answer history during evaluation, which is not realistic, whereas all the others are evaluated with predicted answers. All models are trained with the same protocol, using gold answers as the conversation history for the first half of training epochs (Step 1).

Gold This model uses the ground-truth answers during training and evaluation, thus unrealistic.

No Prediction (No Pred.) This model does not use the predicted answers as the conversation history in either training or evaluation steps.

All Prediction (All Pred.) In contrast to No Pred., this model uses all the predicted answers during both training and evaluation steps.

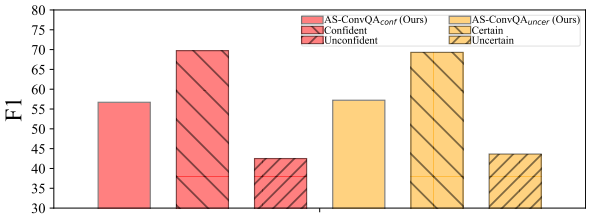

Figure 2: Comparison results of certain (confident) and uncertain (unconfident) predictions on QuAC. Note that a threshold is set as the median of the uncertainty (confidence) values.

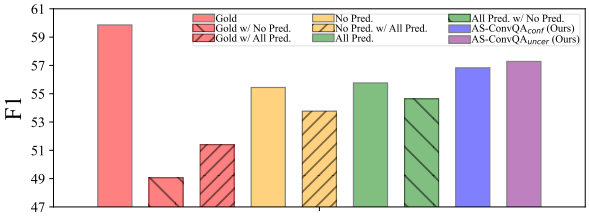

Figure 3: F1-scores on the mismatching evaluation settings for each baseline model on QuAC, either using all of the predictions or none of them as the previous answer history.

CoQAM For training, this model uses the random sampling scheme represented in Equation 4, which samples either predicted or ground-truth answers with coin-flipping (Mandya et al., 2020). For evaluation, it uses all the predictions as the history.

Robust-P Similar to CoQAM, this model uses a heuristic answer sampling scheme in a random manner, but increases the predicted answer sampling rate for training (Siblini et al., 2021). Also, it is evaluated with all predicted answers.

Attentive Selection This model uses the attention mechanism to softly select the relevant answers in the history, following previous work (Qu et al., 2019b; Huang et al., 2019; Chen et al., 2020).

AS-ConvQA$_{\texttt{conf}}$ (Ours) This is our model that filters out unconfident answers via confidence values during training and evaluation, after calibration.

AS-ConvQA$_{\texttt{uncer}}$ (Ours) This is also our model that filters out uncertain answers during training and evaluation, after calibrating uncertainty values.

AS-ConvQA$_{\texttt{combine}}$ (Ours) This model combines our confidence and uncertainty modules, where we use the mean of calibrated confidence and (1-uncertainty) values for filtering out samples.

5 Results and Discussion

In this section, we show overall performances of our proposed method along with detailed analyses.

Main Results As Table 1 shows, the proposed AS-ConvQA models including confidence and uncertainty schemes show significant performance gains over all baselines on two different QA models. Interestingly, No Pred. model, which does not utilize previous answers as the conversation history, shows comparable to or even better performance than the other baseline models based on either exploiting all the predicted answers or randomly sampling them with heuristic ratios. This implies that it is more helpful not to use low-quality predicted answers – unconfident or uncertain – at all than to use them. On the other hand, our models take advantage of filtering out probably invalid predictions, thus achieving improved performance.

Moreover, our AS-ConvQA models outperform the attention-based history selection model (i.e., Attentive Selection). This is because, even though previous answers are all incorrect, the attention scheme should leverage some of them (i.e., the sum of attention scores for previous answers should be 1), which leads the model to answer with an inaccurate history. Meanwhile, AS-ConvQA models can ignore possibly wrong predictions, thus decreasing the risk of being affected by the inaccurate history.

Last, when combining confidence and uncertainty modules, the performance is not much further enhanced. To analyze this, we first measure the number of overlapping questions, where each of the AS-ConvQA$_{\texttt{conf}}$ and AS-ConvQA$_{\texttt{uncer}}$ models predicts with higher confidence or lower uncertainty than its median value. Then, we observe that about 74.82% and 77.12% of the questions overlap on QuAC and CoQA, respectively. This indicates that unconfident and uncertain samples are highly correlated, which are likely to be filtered out by both confidence- and uncertainty-based models. In other words, due to similar effects of AS-ConvQA$_{\texttt{conf}}$ and AS-ConvQA$_{\texttt{uncer}}$ models, the performance of combined models is not much improved.

Unconfident and Uncertain Predictions In order to see whether predictions with low confidence or high uncertainty actually correspond to incorrect answers, we compare the performances between the certain (unconfident) and uncertain (confident) predictions. As Figure 2 shows, low-confident and uncertain samples lead to drastic performance degradation. This result corroborates our hypothesis that a prediction with low confidence or high uncertainty acts as an obstacle in ConvQA tasks.

Impact of Realistic Evaluation Setups To see results in realistic settings – not using ground-truth answers during inference – for the Gold model, we

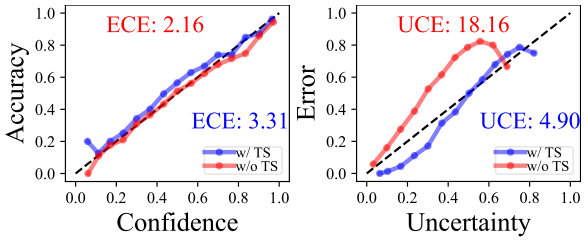

Figure 4: Reliability diagrams with and without temperature scaling (TS), regarding confidence or uncertainty on QuAC.

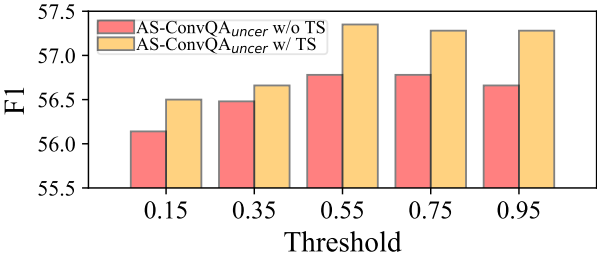

Figure 5: Comparison between the calibrated and not calibrated AS-ConvQA$_{\text{uncer}}$ with varying thresholds on QuAC.

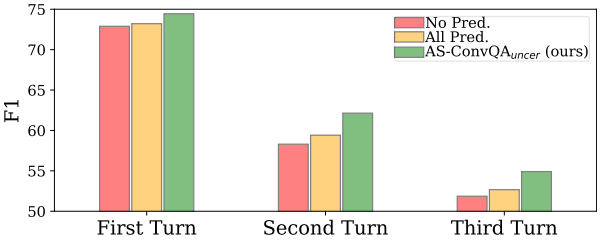

Figure 6: F1-scores for the first, second, and third conversational turns on QuAC with baselines and our AS-ConvQA.

train it with ground-truth answers, and then test either with predicted answers or without them. Figure 3 shows that performances of the Gold model are drastically dropped, and even lower than both No Pred. and All Pred., even though tested on the same strategies. This can be explained with the term of exposure bias (Bengio et al., 2015; Mandya et al., 2020), where a discrepancy exists between training and evaluation, which hinders the model from performing well on test data that differs from training data. Furthermore, this also explains one of the reasons why CoQAM and Robust-P models perform poorly: Since they observe ground-truth answers for training, which are not observable during evaluation, they underperform ours.

Training & Evaluation Discrepancy We have observed a discrepancy between training and evaluation for the Gold model above. Then, the next possible question is whether this discrepancy also happens for models that are trained on the predicted answers, but evaluated in different settings. To see this, we test No Pred. and All Pred. models in a mismatching evaluation setting. As Figure 3 shows, a discrepancy exists for both models, though the gaps are smaller than the Gold model. This implies that even if a ConvQA model is trained on the predictions, the problem of discrepancy should not be ignored. Meanwhile, our proposed models can alleviate such an issue with a selective sampling scheme based on confidence and uncertainty.

Effectiveness of Calibration We show the effect of calibration on confidence and uncertainty values in Figure 4. Regarding confidence, the QA model already generates calibrated scores; thus there is no reason to scale the logit vector with temperature scaling (i.e., w/ temperature scaling yields more errors in terms of ECE). However, regarding uncertainty, the estimated uncertainty scores from the model have high errors in terms of UCE, i.e., not calibrated. Thus, after applying the temperate scaling scheme, the uncertainties become calibrated.

Note that the calibrated uncertainties further contribute to the performance gain, as Figure 5 shows, since the model can observe a broad range of uncertainty values during training, making the model easily capture and reject uncertain predictions.

Effectiveness on Conversational Turns To see how the proposed AS-ConvQA contributes to the quality of the conversation as it proceeds, we further analyze the performances of former and latter conversational turns. As shown in Figure 6, both the former and latter turns benefit from our AS-ConvQA, and the performance improvements are more significant on the latter turn. This result implies that our AS-ConvQA effectively prevents the accumulation of errors, which originate from the incorrect predictions in the previous turns.

Case Study We conduct a case study. As the first example in Table 2 shows, even though both All Pred. and AS-ConvQA$_{\text{uncer}}$ models inaccurately predict $\bar{A}_6$, they handle it differently: All Pred. accepts it, while ours reject it for the subsequent question. In particular, All Pred. model misinterprets 'this' as 'big break' when answering Q_6 as well as Q_7; however, since 'this' actually refers to 'Laputa: Castle in the Sky', All Pred. further propagates the misleading prediction to the next question. By contrast, our model decides not to select $\bar{A}_6$ as a conversational history due to its high uncertainty, thus not repeating the previous mistake when answering Q_7. The proportion of such examples is about 0.52, where our model predicts the previous answer incorrectly but answers the next

Table 2: Examples in a realistic ConvQA evaluation setting for No Pred., All Pred., and AS-ConvQA$_{\text{uncer}}$ (Ours) models.

Case # 1: Our AS-ConvQA removes a previous answer and then predicts a correct answer.

C_1: ... Their collaboration has invited comparisons to the collaborations of Steven Spielberg and John Williams. This big break led to Hisaishi's overwhelming success as a composer of film scores. In 1986, Laputa: Castle in the Sky, would be the first feature to appear under the Studio Ghibli banner, and (A_6) its gentle, faintly melancholic tone would become a familiar trademark of much of the studio's later output. (A_7) And later, in the 1990s, Porco Rosso and Princess Mononoke were released.

Q_7: What other output did the studio release?

	All Pred.	AS-ConvQA$_{\text{uncer}}$ (Ours)
$\mathcal{H}_7$	Q_6: What made this so successful? $\bar{A}_6$: This big break led to Hisaishi's overwhelming success	Q_6: What made this so successful? ~~$\bar{A}_6$: CANNOTANSWER~~ ($s_{uncer} >$ Threshold)
$\bar{A}_7$	In 1986, Laputa: Castle in the Sky, would be the first feature to appear under the Studio Ghibli banner	And later, in the 1990s, Porco Rosso and Princess Mononoke were released.

Case # 2: Our AS-ConvQA keeps a previous answer and then, based on it, predicts a correct answer.

C_2: ... The Walk, Hanson's second studio album with 3CG Records (Fourth overall), was released in the US, Mexico and Canada on July 24. It was released in Japan on February 21 and in the UK on April 30. On May 6, 2007, the 10th anniversary of Hanson Day, (A_5) the band re-recorded their first major label album, Middle Of Nowhere, at The Blank Slate bar in their hometown of Tulsa, Oklahoma. (A_6) The band invited fan club members, causing hundreds to fly to Oklahoma for the acoustic event. Hanson played concerts in the summer of 2007, supporting release of The Walk.

Q_6: Was it well received?

	No Pred.	AS-ConvQA$_{\text{uncer}}$ (Ours)
$\mathcal{H}_6$	Q_5: What did they do on their tenth anniversary? ~~$\bar{A}_5$: the band re-recorded their first major label album, Middle Of Nowhere, at The Blank Slate bar in their hometown of Tulsa, Oklahoma.~~	Q_5: What did they do on their tenth anniversary? $\bar{A}_5$: the band re-recorded their first major label album, Middle Of Nowhere, at The Blank Slate bar in their hometown of Tulsa, Oklahoma. ($s_{uncer} <$ Threshold)
$\bar{A}_6$	CANNOTANSWER	The band invited fan club members, causing hundreds to fly to Oklahoma for the acoustic event.

Case # 3: Our AS-ConvQA predicts an incorrect answer since it filters out a correct previous answer.

C_3: ... Official calendars have also been issued annually from 2004 to 2009, the only exception being 2005. (A_3) Girls Aloud co-wrote an autobiography titled Dreams That Glitter - Our Story. The book, named after a lyric in C̈all the Shots, was published in October 2008 through the Transworld imprint Bantam Press. Before the release, OK! magazine bought the rights to preview and serialise the book. (A_4) In 2007, Girls Aloud signed a PS1.25m one-year deal to endorse hair care brand Sunsilk.

Q_4: What else did they do?

	All Pred.	AS-ConvQA$_{\text{uncer}}$ (Ours)
$\mathcal{H}_4$	Q_3: What else did they do/create? $\bar{A}_3$: Girls Aloud co-wrote an autobiography titled Dreams That Glitter - Our Story.	Q_3: What else did they do/create? ~~$\bar{A}_3$: Girls Aloud co-wrote an autobiography titled Dreams That Glitter - Our Story.~~ ($s_{uncer} >$ Threshold)
$\bar{A}_4$	In 2007, Girls Aloud signed a PS1.25m one-year deal to endorse hair care brand Sunsilk.	Girls Aloud co-wrote an autobiography titled Dreams That Glitter - Our Story.

question with a high F1-score over 50. This emphasizes the importance of our selection scheme, especially when there exist ambiguous words prone to mispredictions.

In addition to this case of removing the uncertain previous prediction, we further compare our model against the No Pred. model in the case where the model predicts with the previous answer history having a low uncertainty value. As the second example in Table 2 shows, while both No Pred. and AS-ConvQA$_{\text{uncer}}$ correctly predict $\bar{A}_5$, No Pred. does not use $\bar{A}_5$ as the answer history when answering the next question, Q_6. However, as $\bar{A}_5$ contains important information of 'it' in Q_6, No Pred. model gives an inaccurate answer to Q_6, since the model is confused about what 'it' refers to. On the other hand, our model selects $\bar{A}_5$ as the answer history due to its low uncertainty value, thereby correctly predicting $\bar{A}_6$ with the previous prediction $\bar{A}_5$. In a third example of Table 2, we show the potential failure of our model, which is discussed in the Limitations section after Section 6.

6 Conclusion

In this work, in order to tackle the challenge of inaccurately predicted answers in the conversation history, we proposed a novel answer selection scheme based on their confidence and uncertainty values. We further calibrated the output values of the model to match the model's predicted confidence and uncertainty to its correct likelihood and error, which makes our answer selection scheme more reliable. The experimental results and analyses demonstrate that AS-ConvQA significantly improves the ConvQA model performance in a realistic evaluation setting without making any architectural changes.

Limitations

While we show the clear advantages of using our AS-ConvQA in realistic ConvQA tasks with both quantitative and qualitative perspectives, there could be possible failures: estimated confidence and uncertainty of a model's prediction do not match its actual correctness. For instance, the third example in Table 2 shows that AS-ConvQA$_{\texttt{uncer}}$ gives an incorrect answer to Q_4, since it removes the correctly predicted previous answer (i.e., $\bar{A}_3$) due to its incorrectly estimated uncertainty. Specifically, both Q_3 and Q_4 ask the additional information: 'What else did they do?'. However, the erroneous deletion of $\bar{A}_3$ makes our model bound to the previous question, repeatedly giving the same answer as $\bar{A}_3$. This implies that AS-ConvQA$_{\texttt{uncer}}$ sometimes assigns high uncertainty to the correct prediction and filters it, which may mislead the model, especially for the one that requires careful attention to the context with the previous answer. Therefore, as future work, one may improve mechanisms to measure incorrectness of predictions.

Ethics Statement

As the need for fully autonomous conversational agents has been rapidly emerging, it is crucial to consider whether ConvQA models can correctly answer a sequence of questions in a realistic setting, in which gold answers for previous questions are unavailable. We note that, in such a challenging setting, our work contributes to the improved performance by selectively using predicted answers with model confidence and uncertainty instead of using predefined gold answers. However, as ConvQA models predict answers based on the given paragraph, we should further consider a scenario where the paragraph itself is not trustworthy, sometimes having offensive contents. Subsequently, this may lead the entire conversation vulnerable to generating unexpected and undesired texts. While this is not the concern raised from our proposed AS-ConvQA models themselves, we still have to make an effort to prevent such an undesirable behavior.

Acknowledgements

This work was supported by Institute for Information and communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No. 2018-0-00582, Prediction and augmentation of the credibility distribution via linguistic analysis and automated evidence document collection).

References

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul W. Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Inf. Fusion*, 76:243–297.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *NeurIPS*, pages 1171–1179.

Yu Chen, Lingfei Wu, and Mohammed J. Zaki. 2020. Graphflow: Exploiting conversation flow with graph neural networks for conversational machine comprehension. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 1230–1236. ijcai.org.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2174–2184. Association for Computational Linguistics.

Soham Dan and Dan Roth. 2021. On the effects of transformer size on in- and out-of-domain calibration. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 2096–2101. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Ahmed Elgohary, Denis Peskov, and Jordan L. Boyd-Graber. 2019. Can you unpack that? learning to rewrite questions-in-context. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5917–5923. Association for Computational Linguistics.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24,*

2016, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 1050–1059. JMLR.org.

Devin Guillory, Vaishaal Shankar, Sayna Ebrahimi, Trevor Darrell, and Ludwig Schmidt. 2021. Predicting with confidence on unseen distributions. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 1114–1124. IEEE.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Sebastian Houben, Stephanie Abrecht, Maram Akila, Andreas Bär, Felix Brockherde, Patrick Feifel, Tim Fingscheidt, Sujan Sai Gannamaneni, Seyed Eghbal Ghobadi, Ahmed Hammam, Anselm Haselhoff, Felix Hauser, Christian Heinzemann, Marco Hoffmann, Nikhil Kapoor, Falk Kappel, Marvin Klingner, Jan Kronenberger, Fabian Küppers, Jonas Löhdefink, Michael Mlynarski, Michael Mock, Firas Mualla, Svetlana Pavlitskaya, Maximilian Poretschkin, Alexander Pohl, Varun Ravi-Kumar, Julia Rosenzweig, Matthias Rottmann, Stefan Rüping, Timo Sämann, Jan David Schneider, Elena Schulz, Gesina Schwalbe, Joachim Sicking, Toshika Srivastava, Serin Varghese, Michael Weber, Sebastian Wirkert, Tim Wirtz, and Matthias Woehrle. 2022. *Inspect, Understand, Overcome: A Survey of Practical Methods for AI Safety*, pages 3–78. Springer International Publishing, Cham.

Hsin-Yuan Huang, Eunsol Choi, and Wen-tau Yih. 2019. Flowqa: Grasping flow in history for conversational machine comprehension. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in bayesian deep learning for computer vision? In *NeurIPS*, pages 5574–5584.

Gangwoo Kim, Hyunjae Kim, Jungsoo Park, and Jaewoo Kang. 2021. Learn to resolve conversational dependency: A consistency training framework for conversational question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6130–6141. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Max-Heinrich Laves, Sontje Ihler, Karl-Philipp Kortmann, and Tobias Ortmaier. 2019. Well-calibrated model uncertainty with temperature scaling for dropout variational inference. *ArXiv*, abs/1909.13550.

Huihan Li, Tianyu Gao, Manan Goenka, and Danqi Chen. 2022. Ditch the gold standard: Re-evaluating conversational question answering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8074–8085. Association for Computational Linguistics.

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

Andrey Malinin and Mark J. F. Gales. 2021. Uncertainty estimation in autoregressive structured prediction. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

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 *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 2017–2025. European Language Resources Association.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pages 8024–8035. Curran Associates, Inc.

Gabriel Pereyra, George Tucker, Jan Chorowski, Lukasz Kaiser, and Geoffrey E. Hinton. 2017. Regularizing neural networks by penalizing confident output distributions. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net.

Minghui Qiu, Xinjing Huang, Cen Chen, Feng Ji, Chen Qu, Wei Wei, Jun Huang, and Yin Zhang. 2021. Reinforced history backtracking for conversational question answering. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13718–13726. AAAI Press.

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, SIGIR 2019, Paris, France, July 21-25, 2019*, pages 1133–1136. ACM.

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 *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, pages 1391–1400. ACM.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2383–2392. The Association for Computational Linguistics.

Gonçalo Raposo, Rui Ribeiro, Bruno Martins, and Luísa Coheur. 2022. Question rewriting? assessing its importance for conversational question answering. In *Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part II*, volume 13186 of *Lecture Notes in Computer Science*, pages 199–206. Springer.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. Coqa: A conversational question answering challenge. *Trans. Assoc. Comput. Linguistics*, 7:249–266.

Artem Shelmanov, Dmitry Puzyrev, Lyubov Kupriyanova, Denis Belyakov, Daniil Larionov, Nikita Khromov, Olga Kozlova, Ekaterina Artemova, Dmitry V. Dylov, and Alexander Panchenko. 2021. Active learning for sequence tagging with deep pre-trained models and bayesian uncertainty estimates. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 1698–1712. Association for Computational Linguistics.

Wissam Siblini, Baris Sayil, and Yacine Kessaci. 2021. Towards a more robust evaluation for conversational question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 2: Short Papers), Virtual Event, August 1-6, 2021*, pages 1028–1034. Association for Computational Linguistics.

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2818–2826. IEEE Computer Society.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. Newsqa: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP, Rep4NLP@ACL 2017, Vancouver, Canada, August 3, 2017*, pages 191–200. Association for Computational Linguistics.

Svitlana Vakulenko, Shayne Longpre, Zhucheng Tu, and Raviteja Anantha. 2021. Question rewriting for conversational question answering. In *WSDM '21, The Fourteenth ACM International Conference on Web Search and Data Mining, Virtual Event, Israel, March 8-12, 2021*, pages 355–363. ACM.

Artem Vazhentsev, Gleb Kuzmin, Artem Shelmanov, Akim Tsvigun, Evgenii Tsymbalov, Kirill Fedyanin, Maxim Panov, Alexander Panchenko, Gleb Gusev, Mikhail Burtsev, Manvel Avetisian, and Leonid Zhukov. 2022. Uncertainty estimation of transformer predictions for misclassification detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8237–8252. Association for Computational Linguistics.

Xiao Wang, Hongrui Liu, Chuan Shi, and Cheng Yang. 2021. Be confident! towards trustworthy graph neural networks via confidence calibration. In *NeurIPS*, pages 23768–23779.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Minghao Wu, Yitong Li, Meng Zhang, Liangyou Li, Gholamreza Haffari, and Qun Liu. 2021. Uncertainty-aware balancing for multilingual and multi-domain neural machine translation training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 7291–7305. 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, ICML 2021, 18-24 July 2021, Virtual Event, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

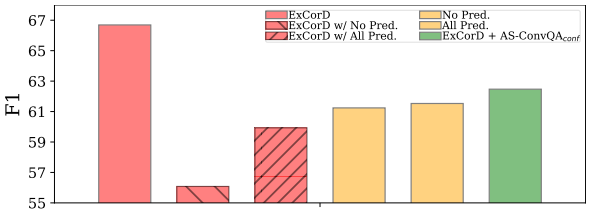

Figure 7: F1-scores on the mismatching evaluation settings for the recent ExCorD model (Kim et al., 2021) on QuAC.

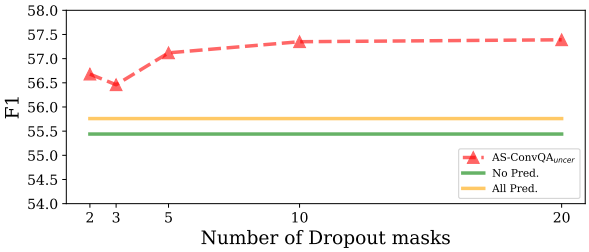

Figure 8: F1 scores with varying dropout numbers on QuAC.

A Experimental Implementation Details

We implement all models using PyTorch (Paszke et al., 2019) and Transformers library (Wolf et al., 2020). For language models, we use BERT-base and RoBERTa-base models with 110M and 125M parameters, respectively. For training, we set the training epoch as 2 with the batch size of 12, where the first epoch is used for Step 1 while the second epoch is used for Step 2. Furthermore, we optimize all models with the Adam optimizer (Kingma and Ba, 2015) with a learning rate of 3e-5. For computiting resources, we use a single GeForce RTX 3090 GPU with 24GB memory, on which each training epoch requires approximately 4 hours.

For hyperparameters, we search the temperature value τ for temperature scaling with a validation set, in the range of $(0, 2]$. Also, we set the filtering threshold for Step 2, in the range of $[\texttt{median} - 0.25, \texttt{median} + 0.25]$, where median is the median value of confidence or uncertainty for all samples. For the number of dropout masks (i.e., N for the uncertainty estimation in Section 3.3) for measuring uncertainty, we set it as 10.

We use two benchmark ConvQA datasets, which are QuAC[3] (Choi et al., 2018) and CoQA[4] (Reddy et al., 2019). Note that, while our main focus is on predicting the extractive answers within a given context, CoQA is designed for answering question in a free-form text, which might not appear in a given context. Therefore, following the experimental setting from Reddy et al. (2019), we convert the CoQA dataset to our extractive ConvQA setting. In particular, we assume the gold answer as the provided rationale, and then make prediction on it, except for simple yes or no questions. For the yes or no questions, we additionally augment yes and no tokens at the end of the paragraph.

B Additional Experimental Results

Realistic Evaluation of ExCorD Even though we validate a negative impact of exposure bias in Figure 3, we further explore the performance of the unrealistic state-of-the-art model, ExCorD (Kim et al., 2021), that uses gold answer histories, in Figure 7 with the realistic ConvQA setting that uses predicted answers. We observe that, similar to the mismatching evaluation experiments reported in Figure 3, the F1-scores of ExCorD drastically drop when evaluated with No Pred. and All Pred. settings, which aligns with our motivation. On the other hand, the performance is much improved by further adapting our AS-ConvQA on ExCorD. This result indicates the importance of filtering unnecessary predictions together with the applicability of our AS-ConvQA model in a realistic setting.

Varying the Number of Dropout Masks In order to understand how the number of dropout masks (i.e., N used for uncertainty estimation in Section 3.3) affects the performance, we vary the number of masks for AS-ConvQA$_{\texttt{uncer}}$. As Figure 8 shows, the performance is stabilized after a certain number of masks (i.e., 5). This indicates the importance of setting an appropriate sampling number, since approximating the uncertainty with a small number of masks is likely to be inaccurate.

[3]https://quac.ai/

[4]https://stanfordnlp.github.io/coqa/