On the Robustness of Dialogue History Representation in Conversational Question Answering: A Comprehensive Study and a New Prompt-based Method

Zorik Gekhman^{T*} **Nadav Oved**^{T*} **Orgad Keller**^{G} **Idan Szpektor**^{G} **Roi Reichart**^{T}
^{T}Technion - Israel Institute of Technology, Israel ^{G}Google Research, Israel
{zorik@campus.|nadavo@campus.|roiri@}technion.ac.il
{orgad|szpektor}@google.com

Abstract

Most work on modeling the conversation history in Conversational Question Answering (CQA) reports a single main result on a common CQA benchmark. While existing models show impressive results on CQA leaderboards, it remains unclear whether they are robust to shifts in setting (sometimes to more realistic ones), training data size (e.g., from large to small sets) and domain. In this work, we design and conduct the first large-scale robustness study of history modeling approaches for CQA. We find that high benchmark scores do not necessarily translate to strong robustness, and that various methods can perform extremely differently under different settings. Equipped with the insights from our study, we design a novel prompt-based history modeling approach and demonstrate its strong robustness across various settings. Our approach is inspired by existing methods that highlight historic answers in the passage. However, instead of highlighting by modifying the passage token embeddings, we add textual prompts directly in the passage text. Our approach is simple, easy to plug into practically any model, and highly effective, thus we recommend it as a starting point for future model developers. We also hope that our study and insights will raise awareness to the importance of robustness-focused evaluation, in addition to obtaining high leaderboard scores, leading to better CQA systems.[1]

1 Introduction

Conversational Question Answering (CQA) involves a dialogue between a user who asks questions and an agent that answers them based on a given document. CQA is an extension of the traditional single-turn QA task (Rajpurkar et al., 2016), with the major difference being the presence of the conversation history, which requires effective *history modeling* (Gupta et al., 2020). Previous work demonstrated that the straightforward approach of concatenating the conversation turns to the input is lacking (Qu et al., 2019a), leading to various proposals of architecture components that explicitly model the conversation history (Choi et al., 2018; Huang et al., 2019; Yeh and Chen, 2019; Qu et al., 2019a,b; Chen et al., 2020; Kim et al., 2021). However, there is no single agreed-upon setting for evaluating the effectiveness of such methods, with the majority of prior work reporting a single main result on a CQA benchmark, such as CoQA (Reddy et al., 2019) or QuAC (Choi et al., 2018).

While recent CQA models show impressive results on these benchmarks, such a single-score evaluation scheme overlooks aspects that can be essential in real-world use-cases. First, QuAC and CoQA contain large annotated training sets, which makes it unclear whether existing methods can remain effective in small-data settings, where the annotation budget is limited. In addition, the evaluation is done in-domain, ignoring the model's robustness to domain shifts, with target domains that may even be unknown at model training time. Furthermore, the models are trained and evaluated using a ''clean'' conversation history between 2 humans, while in reality the history can be ''noisy'' and less fluent, due to the incorrect answers by the model (Li et al., 2022). Finally, these benchmarks mix the impact of advances in pre-trained language models (LMs) with conversation history modeling effectiveness.

In this work, we investigate the *robustness* of *history modeling* approaches in CQA. We ask whether high performance on existing benchmarks also indicates strong robustness. To address this

[*] Authors contributed equally to this work.

[1] Our code and data are available at: `https://github.com/zorikg/MarCQAp`.

Transactions of the Association for Computational Linguistics, vol. 11, pp. 351–366, 2023. https://doi.org/10.1162/tacl_a_00549
Action Editor: Preslav I. Nakov. Submission batch: 7/2022; Revision batch: 11/2022; Published 4/2023.

	Training	In-Domain Evaluation		Out-Of-Domain Evaluation							
Data source	QuAC	QuAC	QuAC-NH	CoQA					DoQA		
Domain				children stories	literature	mid-high school	news	wikipedia	cooking	movies	travel
# Examples	83,568	7,354	10,515	1,425	1,630	1,653	1,649	1,626	1,797	1,884	1,713
# Conversations	11,567	1,000	1,204	100	100	100	100	100	400	400	400

Table 1: Datasets statistics.

question, we carry out the first large-scale robustness study using 6 common modeling approaches. We design 5 robustness-focused evaluation settings, which we curate based on 4 existing CQA datasets. Our settings are designed to evaluate efficiency in low-data scenarios, the ability to scale in a high-resource setting, as well as robustness to domain-shift and to noisy conversation history. We then perform a comprehensive robustness study, where we evaluate the considered methods in our settings.

We focus exclusively on *history modeling*, as it is considered the most significant aspect of CQA (Gupta et al., 2020), differentiating it from the classic single-turn QA task. To better reflect the contribution of the history modeling component, we adapt the existing evaluation metric. First, to avoid differences which stem from the use of different pre-trained LMs, we fix the underlying LM for all the evaluated methods, re-implementing all of them. Second, instead of focusing on final scores on a benchmark, we focus on each model's improvement ($\Delta\%$) compared to a baseline QA model that has no access to the conversation history.

Our results show that history modeling methods perform very differently in different settings, and that approaches that achieve high benchmark scores are not necessarily robust under low-data and domain-shift settings. Furthermore, we notice that approaches that highlight historic answers within the document by modifying the document embeddings achieve the top benchmark scores, but their performance is surprisingly lacking in low-data and domain-shift settings. We hypothesize that history highlighting yields high-quality representation, but since the existing highlighting methods add dedicated embedding parameters, specifically designed to highlight the document's tokens, they are prone to over-fitting.

These findings motivate us to search for an alternative history modeling approach with improved robustness across different settings. Following latest trends w.r.t. prompting in NLP (Liu et al., 2021) we design MarCQAp, a novel prompt-based approach for history modeling, which adds textual prompts within the grounding document in order to highlight previous answers from the conversation history. While our approach is inspired by the embedding-based highlighting methods, it is not only simpler, but it also shows superior *robustness* compared to other evaluated approaches. As MarCQAp is prompt-based, it can be easily combined with any architecture, allowing to fine-tune any model with a QA architecture for the CQA task with minimal effort. Thus, we hope that it will be adopted by the community as a useful starting point, owing to its simplicity, as well as high effectiveness and robustness. We also hope that our study and insights will encourage more robustness-focused evaluations, in addition to obtaining high leaderboard scores, leading to better CQA systems.

2 Preliminaries

2.1 CQA Task Definition and Notations

Given a text passage P, the current question q_k and a conversation history $\mathcal{H}_k$ in a form of a sequence of previous questions and answers $\mathcal{H}_k = (q_1, a_1, \ldots, q_{k-1}, a_{k-1})$, a CQA model predicts the answer a_k based on P as a knowledge source. The answers can be either spans within the passage P (*extractive*) or free-form text (*abstractive*).

2.2 CQA Datasets

Full datasets statistics are presented in Table 1.

QuAC (Choi et al., 2018) and **CoQA** (Reddy et al., 2019) are the two leading CQA datasets, with different properties. In QuAC, the questions are more exploratory and open-ended with longer answers that are more likely to be followed up. This makes QuAC more challenging and realistic.

We follow the common practice in recent work (Qu et al., 2019a,b; Kim et al., 2021; Li et al., 2022), focusing on QuAC as our main dataset, using its training set for training and its validation set for in-domain evaluation (the test set is hidden, reserved for a leaderboard challenge).

We use CoQA for additional pre-training or for domain-shift evaluation.

DoQA (Campos et al., 2020) is another CQA dataset with dialogues from the Stack Exchange online forum. Due to its relatively small size, it is typically used for testing transfer and zero-shot learning. We use it for domain-shift evaluation.

QuAC Noisy-History (QuAC-NH) is based on a datatset of human-machine conversations collected by Li et al. (2022), using 100 passages from the QuAC validation set. While Li et al. used it for human evaluation, we use it for automatic evaluation, leveraging the fact that the answers are labeled for correctness, which allows us to use the *correct* answers as labels.

In existing CQA datasets, each conversation $(q_1, a_1, .., q_m, a_m)$ and the corresponding passage P, are used to create m examples $\{E_k\}_{k=1}^m = \{(P, \mathcal{H}_k, q_k)\}_{k=1}^m$, where $\mathcal{H}_k = (q_1, a_1, \ldots q_{k-1}, a_{k-1})$. a_k is then used as a label for E_k. Since QuAC-NH contains incorrect answers, if a_k is incorrect we discard E_k to avoid corrupting the evaluation set with incorrectly labeled examples. We also filtered out invalid questions (Li et al., 2022) and answers that did not appear in P.[2]

2.3 CQA Related Work

Conversation History Modeling is the major challenge in CQA (Gupta et al., 2020). Early work used recurrent neural networks (RNNs) and variants of attention mechanisms (Reddy et al., 2019; Choi et al., 2018; Zhu et al., 2018). Another trend was to use flow-based approaches, which generate a latent representation for the tokens in $\mathcal{H}_k$, using tokens from P (Huang et al., 2019; Yeh and Chen, 2019; Chen et al., 2020). Modern approaches, which are the focus of our work, leverage Transformer-based (Vaswani et al., 2017) pre-trained language models.

The simplest approach to model the history with pre-trained LMs is to concatenate $\mathcal{H}_k$ with q_k and P (Choi et al., 2018; Zhao et al., 2021). Alternative approaches rewrite q_k based on $\mathcal{H}_k$ and use the rewritten questions instead of $\mathcal{H}_k$ and q_k (Vakulenko et al., 2021), or as an additional training signal (Kim et al., 2021). Another fundamental approach is to highlight historic answers within P by modifying the passage's token embeddings (Qu et al., 2019a,b). Qu et al. also introduced a component that performs dynamic history selection after each turn is encoded. Yet, in our corresponding baseline we utilize only the historic answer highlighting mechanism, owing to its simplicity and high effectiveness. A contemporaneous work proposed a global history attention component, designed to capture long-distance dependencies between conversation turns (Qian et al., 2022).[3]

	Pre-trained LM Size	Training	Evaluation
Standard	Base	QuAC	QuAC
High-Resource	Large	CoQA + QuAC	QuAC
Low-Resource	Base	QuAC smaller samples	QuAC
Domain-Shift	Base	QuAC	CoQA + DoQA
Noisy-History	Base	QuAC	QuAC-NH

Table 2: Summary of our proposed settings.

3 History Modeling Study

In this work, we examine the effect of a model's history representation on its robustness. To this end, we evaluate different approaches under several settings that diverge from the *standard* supervised benchmark (§3.1). This allows us to examine whether the performance of some methods deteriorates more quickly than others in different scenarios. To better isolate the gains from history modeling, we measure performance compared to a baseline QA model which has no access to $\mathcal{H}_k$ (§3.2), and re-implement all the considered methods using the same underlying pre-trained language model (LM) for text representation (§3.3).

3.1 Robustness Study Settings

We next describe each comparative setting in our study and the rationale behind it, as summarized in Table 2. Table 1 depicts the utilized datasets.

Standard. Defined by Choi et al. (2018), this setting is followed by most studies. We use a medium-sized pre-trained LM for each method, commonly known as its *base* version, then fine-tune and evaluate the models on QuAC.

High-Resource. This setting examines the extent to which methods can improve their performance when given more resources. To this

[2] Even though Li et al. only used extractive models, a small portion of the answers did not appear in the passage.

[3] Published 2 weeks before our submission.

end, we use a *large* pre-trained LM, perform additional pre-training on CoQA (with the CQA objective), and then fine-tune and evaluate on QuAC.

Low-Resource. In this setting, we examine the resource efficiency of the history modeling approaches by reducing the size of the training set. This setting is similar to the *standard* setting, except that we fine-tune on smaller samples of QuAC's training set. For each evaluated method we train 4 model variants: 20%, 10%, 5%, and 1%, reflecting the percentage of training data retained.

Domain-Shift. This setting examines *robustness* to domain shift. To this end, we use the 8 domains in the CoQA and DoQA datasets as test sets from unseen target domains, evaluating the models trained under the *standard* setting on these test-sets.

Noisy-History. This setting examines robustness to noisy conversation history, where the answers are sometimes incorrect and the conversation flow is less fluent. To this end, we evaluate the models trained under the *standard* setting on the QuAC-NH dataset, consisting of conversations between humans and *other* CQA models (§2.2). We note that a full human-machine evaluation requires a human in the loop. We choose to evaluate against *other* models predictions as a middle ground. This allows us to test the models' behavior on noisy conversations with incorrect answers and less fluent flow, but without a human in the loop.

3.2 Evaluation Metric

The standard CQA evaluation metric is the average word-level F1 score (Rajpurkar et al., 2016; Choi et al., 2018; Reddy et al., 2019; Campos et al., 2020).[4] Since we focus on the impact of history modeling, we propose to consider each model's improvement in F1 ($\Delta\%$) compared to a baseline QA model that has no access to the dialogue history.

3.3 Pre-trained LM

To control for differences which stem from the use of different pre-trained LMs, we re-implement all the considered methods using the Longformer (Beltagy et al., 2020), a sparse-attention Transformer designed to process long input sequences. It is therefore a good fit for handling the conversation history and the source passage as a combined (long) input. Prior work usually utilized dense-attention Transformers, whose input length limitation forced them to truncate $\mathcal{H}_k$ and split P into chunks, processing them separately and combining the results (Choi et al., 2018; Qu et al., 2019a,b; Kim et al., 2021; Zhao et al., 2021). This introduces additional complexity and diversity in the implementation, while with the Longformer we can keep implementation simple, as this model can attend to the entire history and passage.

We would also like to highlight RoR (Zhao et al., 2021), which enhances a dense-attention Transformer to better handle long sequences. Notably, the state-of-the-art result on QuAC was reported using ELECTRA+RoR with simple history concatenation (see CONCAT in §3.4). While this suggests that ELECTRA+RoR can outperform the Longformer, since our primary focus is on analyzing the robustness of different history modeling techniques rather than on long sequence modeling, we opt for a general-purpose commonly used LM for long sequences, which exhibits competitive performance.

3.4 Evaluated Methods

In our study we choose to focus on modern history modeling approaches that leverage pre-trained LMs. These models have demonstrated significant progress in recent years (§2.3).

NO HISTORY A classic single-turn QA model without access to $\mathcal{H}_k$. We trained a Longformer for QA (Beltagy et al., 2020), using q_k and P as a single packed input sequence (ignoring $\mathcal{H}_k$). The model then *extracts* the answer span by predicting its start and end positions within P.

In contrast to the rest of the evaluated methods, we do not consider this method as a baseline for history modeling, but rather as a reference for calculating our $\Delta\%$ metric. As discussed in §3.2, we evaluate all history modeling methods for their ability to improve over this model.

CONCAT Concatenating $\mathcal{H}_k$ to the input (i.e., to q_k and P), which is (arguably) the most straightforward way to model the history (Choi et al., 2018; Qu et al., 2019a; Zhao et al., 2021). Other than the change to the input, this model architecture and training is identical to NO HISTORY.

[4] We follow the calculation presented in Choi et al. (2018).

Rewrite This approach was proposed in Vakulenko et al. (2021). It consists of a pipeline of two models, question rewriting (QR) and question answering (QA). An external QR model first generates a rewritten question $\tilde{q}_k$, based on q_k and $\mathcal{H}_k$. $\tilde{q}_k$ and P are then used as input to a standard QA model, identical to No History, but trained with the rewritten questions. For the external QR model we follow Lin et al. (2020), Vakulenko et al. (2021), and Kim et al. (2021) and fine-tune T5-base (Raffel et al., 2020) on the CANARD dataset (Elgohary et al., 2019). We use the same QR model across all the settings in our study (§3.1), meaning that in the *low-resource* setting we limit only the CQA data, which is used to train the QA model.

Rewrite$_C$ Hypothesizing that there is useful information in $\mathcal{H}_k$ on top of the rewritten question $\tilde{q}_k$, we combine Rewrite and Concat, obtaining a model which is similar to Concat, except that it replaces q_k with $\tilde{q}_k$.

ExCorD$_{LF}$ Our implementation of the ExCorD approach, proposed in Kim et al. (2021). Instead of rewriting the original question, q_k, at inference time (Rewrite), ExCorD uses the rewritten question only at training time as a regularization signal when encoding the original question.

HAE$_{LF}$ Our implementation of the HAE approach proposed in Qu et al. (2019a), which highlights the conversation history within P. Instead of concatenating $\mathcal{H}_k$ to the input, HAE highlights the historic answers $\{a_i\}_{i=1}^{k-1}$ within P, by modifying the passage token embeddings. HAE adds an additional dedicated embedding layer with 2 learned embedding vectors, denoting whether a token from P appears in any historic answers or not.

PosHAE$_{LF}$ Our implementation of the PosHAE approach proposed in Qu et al. (2019b), which extends HAE by adding positional information. The embedding matrix is extended to contain a vector per conversation turn, each vector representing the turn that the corresponding token appeared in.

3.5 Implementation Details

We fine-tune all models on QuAC for 10 epochs, employ an accumulated batch size of 640, a weight decay of 0.01, and a learning rate of $3 \cdot 10^{-5}$. In the high-resource setup, we also pre-train on CoQA for 5 epochs. We use a maximum output length of 64 tokens. Following Beltagy et al. (2020), we set Longformer's global attention to all the tokens of q_k. We use the cross-entropy loss and *AdamW* optimizer (Kingma and Ba, 2015; Loshchilov and Hutter, 2019). Our implementation makes use of the *HuggingFace Transformers* (Wolf et al., 2020), and *PyTorch-Lightning* libraries.[5]

For the *base* LM (used in all settings except *high-resource*) we found that a Longformer that was further pre-trained on SQuADv2 (Rajpurkar et al., 2018),[6] achieved consistently better performance than the base Longformer. Thus, we adopted it as our *base* LM. For the *large* LM (used in the *high-resource* setting) we used Longformer-large.[7]

In §5, we introduce a novel method (MarCQAp) and perform statistical significance tests (Dror et al., 2018, 2020). Following Qu et al. (2019b), we use the Student's paired t-test with $p < 0.05$, to compare MarCQAp to all other methods in each setting.

In our re-implementation of the evaluated methods, we carefully followed their descriptions and implementation details as published by the authors in their corresponding papers and codebases. A key difference in our implementation is the use of a long sequence Transformer, which removes the need to truncate $\mathcal{H}_k$ and split P into chunks (§3.3). This simplifies our implementation and avoids differences between methods.[8] Table 3 compares between our results and those reported in previous works. In almost all cases we achieved

	Original Work	Original LM	Original Result	Our Impl.
Concat	Qu et al. (2019a)	BERT	62.0	65.8
Rewrite	Vakulenko et al. (2021)	BERT	Not Reported	64.6
Rewrite$_C$	N/A (this baseline was first proposed in this work)			67.3
ExCorD	Kim et al. (2021)	RoBERTa	67.7	67.5
HAE	Qu et al. (2019a)	BERT	63.9	68.9
PosHAE	Qu et al. (2019b)	BERT	64.7	69.8

Table 3: F1 scores comparison between original implementations and ours (using Longformer as the LM), for all methods described in §3.4, in the *standard* setting.

[5] https://github.com/PyTorchLightning/pytorch-lightning.

[6] https://huggingface.co/mrm8488/longformer-base-4096-finetuned-squadv2.

[7] https://huggingface.co/allenai/longformer-large-4096.

[8] The maximum length limit of $\mathcal{H}_k$ varies between different works, as well as how sub-document chunks are handled.

Setting	*Low-Resource*					*Standard*	*High-Resource*
LM	Longformer-base Pre-trained SQuAD					Longformer-base Pre-trained SQuAD	Longformer-large Pre-trained CoQA
Training set size	800 (1%)	$4K$ (5%)	$8K$ (10%)	$16K$ (20%)	Avg Δ%	$80K$ (100%)	$80K$ (100%)
No History	45.0	50.0	52.9	55.4	–	60.4	65.6
concat	43.9 (-2.4%)	51.2 (+2.4%)	53.4 (+0.9%)	57.8 (+4.3%)	+1.3%	65.8 (+8.9%)	72.3 (+10.2%)
Rewrite	46.5 (+3.3%)	54.0 (+8.0%)	56.4 (+6.6%)	59.2 (+6.9%)	+6.2%	64.6 (+7.0%)	69.0 (+5.2%)
Rewrite$_C$	42.3 (-6.0%)	54.4 (+8.8%)	57.2 (+8.1%)	60.6 (+9.4%)	+5.1%	67.3 (+11.4%)	72.5 (+10.5%)
ExCorD$_{LF}$	46.0 (+2.2%)	53.0 (+6.0%)	57.2 (+8.1%)	60.3 (+8.8%)	+6.3%	67.5 (+11.8%)	73.8 (+12.3%)
HAE$_{LF}$	44.5 (-1.1%)	50.8 (+1.6%)	55.0 (+4.0%)	59.8 (+7.9%)	+3.1%	69.0 (+14.2%)	73.2 (+11.4%)
PosHAE$_{LF}$	40.5 (-10.0%)	51.0 (+2.0%)	55.1 (+4.2%)	60.9 (+9.9%)	+1.5%	69.8 (+15.6%)	74.2 (+12.9%)
MarCQAp (§5)	**48.2 (+7.1%)**	**57.4 (+14.8%)**	**61.3 (+15.9%)**	**64.6 (+16.6%)**	**+13.6%**	**70.2 (+16.2%)**	**74.7 (+13.7%)**

Table 4: In-domain F1 and Δ% scores on the full QuAC validation set, for the *standard*, *high-resource* and *low-resource* settings. We color coded the Δ% for positive and negative numbers.

a higher score (probably since Longformer outperforms BERT), with the exception of ExCorD, where we achieved a comparable score (probably since Longformer is actually initialized using RoBERTa's weights [Beltagy et al., 2020]).

4 Results and Analysis

We next discuss the takeaways from our study, where we evaluated the considered methods across the proposed settings. Table 4 presents the results of the *standard*, *high-resource*, and *low-resource* settings. Table 5 further presents the *domain-shift* results. Finally, Table 6 depicts the results of the *noisy-history* setting. Each method is compared to No History by calculating the Δ% (§3.2). The tables also present the results of our method, termed MarCQAp, which is discussed in §5.

We further analyze the effect of the conversation history length in Figure 1, evaluating models from the *standard* setting with different limits on the history length. For instance, when the limit is 2, we expose the model to up to the 2 most recent turns, by truncating $\mathcal{H}_k$.[9]

Key Findings A key goal of our study is to examine the robustness of history modeling approaches to setting shifts. This research reveals limitations of the single-score benchmark-based evaluation adopted in previous works (§4.1), as such scores are shown to be only weakly correlated with low-resource and domain-shift robustness. Furthermore, keeping in mind that history modeling is a key aspect of CQA, our study also demonstrates the importance of isolating the contribution of the history modeling method from other model components (§4.2). Finally, we discover that while existing history highlighting approaches yield high-quality input representations, their robustness is surprisingly poor. We further analyze the history highlighting results and provide possible explanations for this phenomenon (§4.3). This finding is the key motivation for our proposed method (§5).

4.1 High CQA Benchmark Scores do not Indicate Good Robustness

First, we observe some expected general trends: All methods improve on top of No History, as demonstrated by the positive Δ% in the *standard* setting, showing that all the methods can leverage information from $\mathcal{H}_k$. All methods scale with more training data and a larger model (*high-resource*), and their performances drop significantly when the training data size is reduced (*low-resource*) or when they are presented with noisy history. A performance drop is also observed when evaluating on *domain-shift*, as expected in the zero shot setting.

However, not all methods scale equally well and some deteriorate faster than others. This phenomenon is illustrated in Table 7, where the methods are ranked by their scores in each setting. We observe high instability between settings. For instance, PosHAE$_{LF}$ is top performing in 3 settings but is second worst in 2 others. Rewrite is second best in *low-resource*, but among the last ones in other settings. So is the case with Concat: Second best in *domain-shift* but among the worst ones in others. In addition, while all the methods improve when they are exposed to longer histories (Figure 1), some saturate earlier than others.

[9] We exclude Rewrite, since it utilizes $\mathcal{H}_k$ only in the form of the rewritten question. For Rewrite$_C$, we truncate the concatenated $\mathcal{H}_k$ for the CQA model, while the QR model remains exposed to the entire history.

Setting	*Domain-Shift*								
Domain	CoQA Children Sto.	 Literature	 M/H Sch.	 News	 Wikipedia	DoQA Cooking	 Movies	 Travel	Avg Δ%
No Hist.	54.8	42.6	50.3	50.1	58.2	46.9	45.0	44.0	–
Concat	62.2 (+13.5%)	48.0 (+12.7%)	55.3 (+9.9%)	54.9 (+9.6%)	59.9 (+2.9%)	**54.8** (**+16.8%**)	**52.0** (**+15.6%**)	48.4 (+10%)	+11.4%
Rewrite	60.1 (+9.7%)	47.7 (+12.0%)	55.0 (+9.3%)	54.8 (+9.4%)	60.9 (+4.6%)	44.6 (-4.9%)	43.2 (-4.0%)	40.9 (-7.0%)	+3.6%
Rewrite$_C$	62.7 (+14.4%)	49.0 (+15.0%)	56.7 (+12.7%)	55.2 (+10.2%)	59.4 (+2.1%)	52.0 (+10.9%)	49.1 (+9.1%)	46.4 (+5.5%)	+10.0%
ExCorD$_{LF}$	62.7 (+14.4%)	51.5 (+20.9%)	58.2 (+15.7%)	57.0 (+13.8%)	63.6 (+9.3%)	53.7 (+14.5%)	51.1 (+13.6%)	48.6 (+10.5%)	+14.1%
HAE$_{LF}$	61.8 (+12.8%)	50.5 (+18.5%)	56.6 (+12.5%)	55.4 (+10.6%)	60.9 (+4.6%)	45.0 (-4.1%)	45.1 (+0.2%)	45.1 (+2.5%)	+7.2%
PosHAE$_{LF}$	56.6 (+3.3%)	47.4 (+11.3%)	55.4 (+10.1%)	52.7 (+5.2%)	61.7 (+6.0%)	45.6 (-2.8%)	45.8 (+1.8%)	44.7 (+1.6%)	+4.6%
MarCQAp (§5)	**66.7** (**+21.7%**)	**56.4** (**+32.4%**)	**61.8** (**+22.9%**)	**60.8** (**+21.4%**)	**67.5** (**+16.0%**)	53.3 (+13.6%)	51.8 (+15.1%)	**50.1** (**+13.9%**)	**+19.6%**

Table 5: F1 and Δ% scores for the *domain-shift* setting. We color coded the Δ% for positive and negative numbers.

Setting	*Noisy-History*
No History	49.9
Concat	55.3 (+10.8%)
Rewrite	56.0 (+12.2%)
Rewrite$_C$	58.5 (+17.2%)
ExCorD$_{LF}$	56.8 (+13.8%)
HAE$_{LF}$	57.9 (+16.0%)
PosHAE$_{LF}$	60.1 (+20.4%)
MarCQAp (§5)	**62.3** (**+24.9%**)

Table 6: F1 and Δ% scores for the *noisy-history* setting.

standard	*high-resource*	*low-resource*	*domain-shift*	*noisy-history*
PH (+15.6%)	PH (+12.9%)	Ex (+6.3%)	Ex (+14.1%)	PH (+20.4%)
H (+14.2%)	Ex (+12.3%)	R (+6.2%)	C (+11.4%)	R$_C$ (+17.2%)
Ex (+11.8%)	H (+11.4%)	R$_C$ (+5.1%)	R$_C$ (+10.0%)	H (+16.0%)
R$_C$ (+11.4%)	R$_C$ (+10.5%)	H (+3.1%)	H (+7.2%)	Ex (+13.8%)
C (+8.9%)	C (+10.2%)	PH (+1.5%)	PH (+4.6%)	R (+12.2%)
R (+7.0%)	R (+5.2%)	C (+1.3%)	R (+3.6%)	C (+10.8%)

Table 7: Per setting rankings of the methods evaluated in our study (top is best), excluding MarCQAp. C is Concat, R is Rewrite, R$_C$ is Rewrite$_C$, Ex is ExCorD$_{LF}$, H is HAE$_{LF}$, and PH is PosHAE$_{LF}$.

We conclude that *the winner does not take it all*: There are significant instabilities in methods' performance across settings. This reveals the limitations of the existing single-score benchmark evaluation practice, and calls for more comprehensive robustness-focused evaluation.

4.2 The Contribution of the History Modeling Method should be Isolated

In the *high-resource* setting, No History reaches 65.6 F1, higher than many CQA results reported in previous work (Choi et al., 2018; Qu et al., 2019a,b; Huang et al., 2019). Since it is clearly ignoring the history, this shows that significant improvements can stem from simply using a better LM. Thus comparing between history modeling methods that use different LMs can be misleading.

This is further illustrated with HAE$_{LF}$'s and PosHAE$_{LF}$'s results. The score that Kim et al. reported for ExCorD is higher than Qu et al. reported for HAE and PosHAE. While both authors used a setting equivalent to our *standard* setting, Kim et al. used RoBERTa while Qu et al. used BERT, as their underlying LM. It is therefore unclear whether ExCorD's higher score stems from better history representation or from choosing to use RoBERTa. In our study, HAE$_{LF}$ and PosHAE$_{LF}$ actually outperform ExCorD$_{LF}$ in the *standard* setting. This suggests that these methods can perform better than reported, and demonstrates the importance of controlling for the choice of LM when comparing between history modeling methods.

As can be seen in Figure 1, Concat saturates at 6 turns, which is interesting since Qu et al. (2019a) reported saturation at 1 turn in a BERT-based equivalent. Furthermore, Qu et al. observed a performance degradation with more turns, while we observe stability. These differences probably stem from the history truncation in BERT, due to the input length limitation of dense attention Transformers. This demonstrates the advantages of sparse attention Transformers for history modeling evaluation, since the comparison against Concat can be more ''fair''. This comparison is important, since the usefulness of any method should be established by comparing it to the straight-forward solution, which is Concat in case of history modeling.

We would also like to highlight PosHAE$_{LF}$'s F1 scores in the *noisy-history* (60.1) and the 20% *low-resource* setting (60.9), both lower than the 69.8 F1 in the *standard* setting. *Do these performance drops reflect lower effectiveness in modeling the conversation history?* Here the Δ% comes to the rescue. While the Δ% decreased between the *standard* and the 20% settings

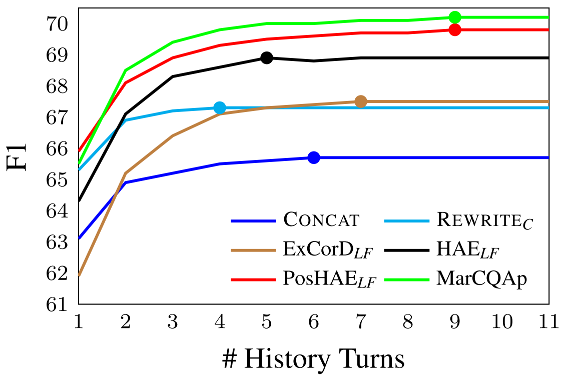


Figure 1: F1 as a function of # history turns, for models from the *standard* setup. The first occurrence of the maximum F1 value (saturation point) is highlighted.

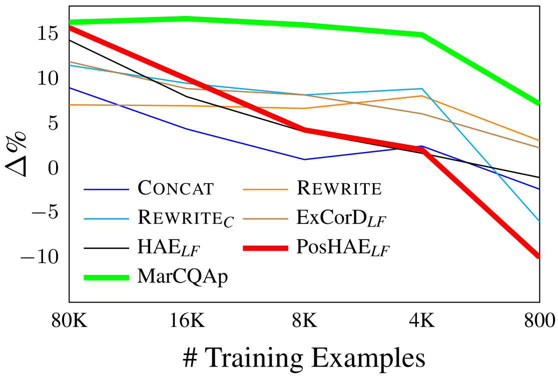


Figure 2: Δ% as a function of # training examples. Results taken from the *standard* and *low-resource* settings.

(15.6 → 9.9), it actually increased in the *noisy-history* setting (to 20.4). This indicates that even though the F1 decreased, the ability to leverage the history actually increased.

We conclude that our study results support the design choices we made, in our effort to better isolate the contribution of the history representation. We recommend future works to compare history modeling methods using the same LM (preferably a long sequence LM), and to measure a Δ% compared to a NO HISTORY baseline.

4.3 History Highlighting is Effective in Resource-rich Setups, but is not Robust

The most interesting results are observed for the history highlighting methods: HAE and PosHAE.

First, when implemented using the Longformer, HAE$_{LF}$ and PosHAE$_{LF}$ perform better than reported in previous work, with 68.9 and 69.8 F1 respectively, compared to 63.9 and 64.7 reported by Qu et al. using BERT. The gap between HAE$_{LF}$ and PosHAE$_{LF}$ demonstrates the effect of the positional information in PosHAE$_{LF}$. This effect is further observed in Figure 1: HAE$_{LF}$ saturates earlier since it cannot distinguish between different conversation turns, which probably yields conflicting information. PosHAE$_{LF}$ saturates at 9 turns, later than the rest of the methods, which indicates that it can better leverage long conversations.

PosHAE$_{LF}$ outperforms all methods in the *standard*, *high-resource*, and *noisy-history* settings,[10] demonstrating the high effectiveness of history highlighting. However, it shows surprisingly poor performance in *low-resource* and *domain-shift* settings, with extremely low average Δ% compared to other methods. The impact of the training set size is further illustrated in Figure 2. We plot the Δ% as a function of the training set size, and specifically highlight PosHAE$_{LF}$ in bold red. Its performance deteriorates significantly faster than others when the training set size is reduced. In the 1% setting it is actually the worst performing method.

This poor robustness could be caused by the additional parameters added in the embedding layer of PosHAE$_{LF}$. Figure 2 demonstrates that properly training these parameters, in order to benefit from this method's full potential, seems to require large amounts of data. Furthermore, the poor *domain-shift* performance indicates that, even with enough training data, this embedding layer seems to be prone to overfitting to the source domain.

We conclude that history highlighting clearly yields a very strong representation, but the additional parameters of the embedding layer seem to require large amounts of data to train properly and over-fit to the source domain. *Is there a way to highlight historic answers in the passage, without adding dedicated embedding layers?*

In §5 we present MarCQAp, a novel history modeling approach that is inspired by PosHAE, adopting the idea of history highlighting. However, instead of modifying the passage embedding, we highlight historic answers by adding textual prompts directly in the input text. By leveraging prompts, we reduce model complexity and remove the need for training dedicated parameters, hoping to mitigate the robustness weaknesses of PosHAE.

[10] We ignore MarCQAp's results in this section.

5 MarCQAp

Motivated by our findings, we design MarCQAp, a novel prompt-based history modeling approach that highlights answers from previous conversation turns by inserting textual prompts in their respective positions within P. By highlighting with prompts instead of embedding vectors, we hope to encode valuable dialogue information, while reducing the learning complexity incurred by the existing embedding-based methods. Thus, we expect MarCQAp to perform well not only in high-resource settings, but also in low-resource and domain adaptation settings, in which prompting methods have shown to be particularly useful (Brown et al., 2020; Le Scao and Rush, 2021; Ben-David et al., 2022).

Prompting often refers to the practice of adding phrases to the input, in order to encourage pre-trained LMs to perform specific tasks (Liu et al., 2021), yet it is also used as a method for injecting task-specific guidance during fine-tuning (Le Scao and Rush, 2021; Ben-David et al., 2022). MarCQAp closely resembles the prompting approach from Ben-David et al. (2022) since our prompts are: (1) discrete (i.e., the prompt is an actual text-string), (2) dynamic (i.e., example-based), and (3) added to the *input* text and the model then makes predictions conditioned on the modified input. Moreover, as in Ben-David et al., in our method the underlying LM is further trained on the downstream task with prompts. However, in contrast to most prompting approaches, which predefine the prompt's location in the input (Liu et al., 2021), our prompts are inserted in different locations for each example. In addition, while most *textual* prompting approaches leverage prompts comprised of natural language, our prompts contain non-verbal symbols (e.g., "*<1>*", see Figure 3 and §5.1), which were proven useful for supervision of NLP tasks. For instance, Aghajanyan et al. (2022) showed the usefulness of structured pre-training by adding HTML symbols to the input text. Finally, to the best of our knowledge, this work is the first to propose a prompting mechanism for the CQA task.

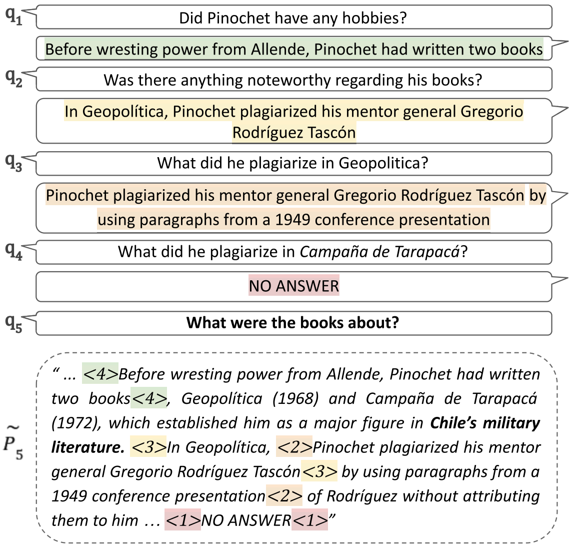

q_1
Did Pinochet have any hobbies?
Before wresting power from Allende, Pinochet had written two books
a_1
q_2
Was there anything noteworthy regarding his books?
In Geopolítica, Pinochet plagiarized his mentor general Gregorio Rodríguez Tascón
a_2
q_3
What did he plagiarize in Geopolitica?
Pinochet plagiarized his mentor general Gregorio Rodríguez Tascón by using paragraphs from a 1949 conference presentation
a_3
q_4
What did he plagiarize in *Campaña de Tarapacá*?
NO ANSWER
a_4
q_5
What were the books about?
$\widetilde{P}_5$
" ... <4>Before wresting power from Allende, Pinochet had written two books<4>, Geopolítica (1968) and Campaña de Tarapacá (1972), which established him as a major figure in ***Chile's military literature.*** *<3>In Geopolítica, <2>Pinochet plagiarized his mentor general Gregorio Rodríguez Tascón<3> by using paragraphs from a 1949 conference presentation<2> of Rodríguez without attributing them to him … <1>NO ANSWER<1>"*

Figure 3: The MarCQAp highlighting scheme: Answers to previous questions are highlighted in the grounding document, which is then provided as input to the model.

5.1 Method

MarCQAp utilizes a standard *single-turn* QA model architecture and input, with the input comprising the current question q_k and the passage P. For each CQA example $(P, \mathcal{H}_k, q_k)$, MarCQAp inserts a textual prompt within P, based on information extracted from the conversation history $\mathcal{H}_k$. In extractive QA, the answer a_k is typically a span within P. Given the input $(P, \mathcal{H}_k, q_k)$, MarCQAp transforms P into an answer-highlighted passage $\widetilde{P}_k$, by constructing a prompt p_k and inserting it within P. p_k is constructed by locating the beginning and end positions of all historic answers $\{a_i\}_{i=1}^{k-1}$ within P, and inserting a unique textual marker for each answer in its respective positions (see example in Figure 3). The input $(\widetilde{P}_k, q_k)$ is then passed to the QA model, instead of (P, q_k).

In abstractive QA, a free form answer is generated based on an evidence span that is first extracted from P. Hence, the final answer does not necessarily appear in P. To support this setting, MarCQAp highlights the historical evidence spans (which appear in P) instead of the generated answers.

To encode positional dialogue information, the markers for $a_j \in \{a_i\}_{i=1}^{k-1}$ include its turn index number in reverse order, that is, $k - 1 - j$. This encodes relative historic positioning w.r.t. the current question q_k, allowing the model to distinguish between the historic answers by their recency.

MarCQAp highlights only the historic answers, since the corresponding questions do not appear in P. While this might lead to information loss, in §5.3 we implement MarCQAp's variants that add the historic questions to the input, and show that the contribution of the historic questions to the performance is minor.[11]

A CQA dialogue may also contain unanswerable questions. Before inserting the prompts, MarCQAp first appends a 'NO ANSWER' string to P.[12] Each historical 'NO ANSWER' is then highlighted with prompts, similarly to ordinary historical answers. For example see a_4 in Figure 3.

MarCQAp has several advantages over prior approaches. First, since it is prompt-based, it does not modify the model architecture, which makes it easier to port across various models, alleviating the need for model-specific implementation and training procedures. Additionally, it naturally represents overlapping answers in P, which was a limitation in prior work (Qu et al., 2019a,b). Overlapping answers contain tokens which relate to multiple turns, yet the existing *token-based* embedding methods encode the relation of a token from P only to a single turn from $\mathcal{H}_k$. Since MarCQAp is *span-based*, it naturally represents overlapping historic answers (e.g., see a_2 and a_3 in Figure 3).

5.2 MarCQAp Evaluation

We evaluate MarCQAp in all our proposed experimental settings (§3.1). As presented in Tables 4, 5, and 6, it outperforms all other methods in all settings. In the *standard*, *high-resource*, and *noisy-history* settings, its performance is very close to PosHAE$_{LF}$,[13] indicating that our prompt-based approach is an effective alternative implementation for the idea of highlighting historical answers. Similarly to PosHAE$_{LF}$, MarCQAp is able to handle long conversations and its performance gains saturate at 9 turns (Figure 1). However, in contrast to PosHAE$_{LF}$, MarCQAp performs especially well in the *low-resource* and the *domain-shift* settings.

In the *low-resource* settings, MarCQAp outperforms all methods by a large margin, with an average Δ% of 13.6% compared to the best baseline with 6.3%. The dramatic improvement over PosHAE$_{LF}$'s average Δ% (1.5% → 13.6%) serves as a strong indication that our prompt-based approach is much more robust. This boost in robustness is best illustrated in Figure 2, which presents the Δ% as a function of the training set size, highlighting PosHAE$_{LF}$ (red) and MarCQAp (green) specifically. An example of MarCQAp's robustness in the *low-resource* setting is provided in Figure 4.

QuAC id: C_721c2ff2b119415c901a3cd1ec2beb28_0

Q1: What was Ronald Ross known for?
A1: *Ronald Ross was noted to be eccentric and egocentric…*
Q2: why was that?
A2: *His professional life appeared to be in constant feud …*
Q3: Any example of what you mean?

Correct Answers:
- *He was openly envious of his mentor Patrick Manson's...*
- *His personal vendetta … became a … tale in science.*

Standard	Concat	NO ANSWER
	Rewrite	*Ronald Ross was noted to be eccentric and egocentric…*
	RewriteC	*His professional life appeared to be in constant feud …*
	ExCorD	*He was openly envious of his mentor Patrick Manson's...*
	HAE	*His personal vendetta … became a … tale in science.*
	PosHAE	*His personal vendetta … became a … tale in science.*
	MarCQAp	*His personal vendetta … became a … tale in science.*
Low Resourse 10%	Concat	NO ANSWER
	Rewrite	*Ronald Ross was noted to be eccentric and egocentric…*
	RewriteC	*His professional life appeared to be in constant feud …*
	ExCorD	*Ronald Ross was noted to be eccentric and egocentric…*
	HAE	NO ANSWER
	PosHAE	NO ANSWER
	MarCQAp	*His personal vendetta … became a … tale in science.*

Figure 4: An example of MarCQAp's robustness in the *low-resource* setting. Even though ExCorD$_{LF}$, HAE$_{LF}$, and PosHAE$_{LF}$ predict correct answers in the *standard* setting, they fail on the same example when the training data size is reduced to 10%. MarCQAp predicts a correct answer in both settings.

In the *domain-shift* settings, MarCQAp is the best performing method in 6 out of 8 domains.[14] On the remaining two domains (Cooking & Movies), CONCAT is the best performing.[15] Notably, MarCQAp's average Δ% (19.6%) is substantially higher compared to the next best method (14.1%). These results serve as additional strong evidence of MarCQAp's robustness.

MarCQAp's Performance Using Different LMs In addition to Longformer, we evaluated MarCQAp using RoBERTa (Liu et al., 2019) and BigBird (Zaheer et al., 2020) in the standard setting. The results are presented in Table 8. MarCQAp shows a consistent positive effect across different LMs, which further highlights its effectiveness.

[11] Which is also in line with the findings in Qu et al. (2019b).

[12] Only if it is not already appended to P, in some datasets the passages are always suffixed with 'NO ANSWER'.

[13] In the *standard* and *high-resource* MarCQAp's improvements over PosHAE$_{LF}$ are not statistically significant.

[14] For the Travel domain MarCQAp's improvement over ExCorD$_{LF}$ is not statistically significant.

[15] The differences between CONCAT and MarCQAp for both domains are not statistically significant.

Model	No History	*MarCQAp*	Δ%
RoBERTa	57.7	68.0	**(+17.9%)**
BigBird	57.6	66.3	(+15.1%)
Longformer$_{base}$	60.0	68.4	(+14.0%)
Longformer$_{squad}$	**60.4**	**70.2**	(+16.6%)

Table 8: MarCQAp's *standard* setting performance across different Transformer-based pre-trained LMs.

We note that since RoBERTa is a dense-attention Transformer with input length limitation of 512 tokens, longer passages are split into chunks. This may lead to some chunks containing part of the historic answers, and therefore partial highlighting by MarCQAp. Our analysis showed that 51% of all examples in QuAC were split into several chunks, and 61% the resulted chunks contained partial highlighting. MarCQAp's strong performance with RoBERTa suggests that it can remain effective even with partial highlighting.

Official QuAC Leaderboard Results For completeness, we submitted our best performing model (from the *high-resource* setting) to the official QuAC leaderboard,[16] evaluating its performance on the hidden test set. Table 9 presents the results.[17] MarCQAp achieves a very competitive score of **74.0** F1, very close to the published state-of-the art (RoR by Zhao et al. [2021] with 74.9 F1), yet with a much simpler model.[18]

5.3 Prompt Design

Recall that MarCQAp inserts prompts at the beginning and end positions for each historical answer within P (Figure 3). The prompts are designed with predefined marker symbols and include the answer's turn index (e.g., "<1>"). This design builds on 3 main assumptions: (1) textual prompts can represent conversation history information, (2) the positioning of the prompts within P facilitates highlighting of historical answers, and (3) indexing the historical answers encodes valuable information. We validate our design assumptions by comparing MarCQAp against ablated variants (Table 10).

BiDAF++ w/ 2-Context (Choi et al., 2018)	60.1
HAE (Qu et al., 2019a)	62.4
FlowQA (Huang et al., 2019)	64.1
GraphFlow (Chen et al., 2020)	64.9
HAM (Qu et al., 2019b)	65.4
FlowDelta (Yeh and Chen, 2019)	65.5
GHR (Qian et al., 2022)	73.7
RoR (Zhao et al., 2021)	**74.9**
MarCQAp (Ours)	74.0

Table 9: Results from the official QuAC leaderboard, presenting F1 scores for the hidden test set, for MarCQAp and other models with published papers.

To validate assumption (1), we compare MarCQAp to MarCQAp$_C$, a variant which adds $\mathcal{H}_k$ to the input, in addition to $\widetilde{P}_k$ and q_k. MarCQAp$_C$ is exposed to information from $\mathcal{H}_k$ via two sources: The concatenated $\mathcal{H}_k$ and the MarCQAp prompt within $\widetilde{P}_k$. We observe a negligible effect,[19] suggesting that MarCQAp indeed encodes information from the conversation history, since providing $\mathcal{H}_k$ does not add useful information on top of $\widetilde{P}_k$.

To validate assumptions (2) and (3), we use two additional MarCQAp's variants. *Answer Pos* inserts a constant predefined symbol ("<>"), in each answer's beginning and end positions within P (i.e., similar to MarCQAp, but without turn indexing). *Random Pos* inserts the same number of symbols but in random positions within P.

Answer Pos achieves a Δ% of 12.7%, while *Random Pos* achieves 1.7%. This demonstrates that the positioning of the prompts within P is crucial, and that most of MarCQAp's performance gains stem from its prompts positioning w.r.t. historical answers $\{a_i\}_{i=1}^{k-1}$. When the prompts are inserted at meaningful positions, the model seems to learn to leverage these positions in order to derive an effective history representation. Surprisingly, *Random Pos* leads to a minor improvement of 1.7%.[20] Finally, MarCQAp's improvement over *Answer Pos* (a Δ% of 15.9% compared to 12.7%), indicates that answer indexing encodes valuable information, helping us validate assumption (3).

Finally, since textual prompts allow for easy injection of additional information, we make

[16] https://quac.ai.

[17] The leaderboard contains additional results for models which (at the time of writing) include no descriptions or published papers, rendering them unsuitable for fair comparison.

[18] See §3.3 for a discussion of RoR.

[19] The difference is not statistically significant.

[20] The difference is statistically significant, we did not further investigate the reasons behind this particular result.

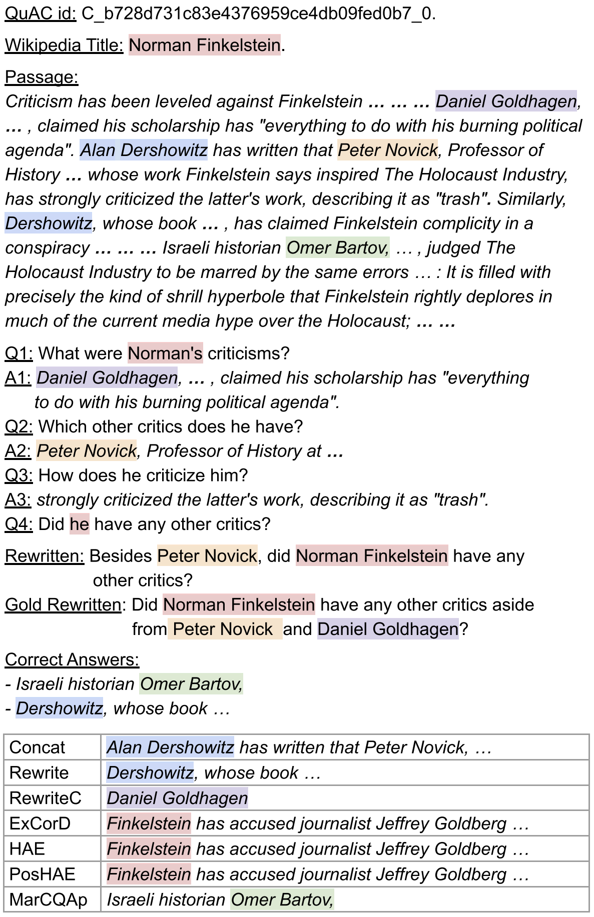
No History	52.9
Random Pos	53.8 (+1.7%)
Answer Pos	59.6 (+12.7%)
Full Q	59.2 (+11.9%)
Word from Q	60.4 (+14.2%)
Word from Q + Index	60.7 (+14.8%)
MarCQAp$_C$	61.5 (+16.3%)
MarCQAp	61.3 (+15.9%)

Table 10: F1 and Δ% scores for MarCQAp's ablated variants, in the 10% setup of the *low-resource* setting.

several initial attempts in this direction, injecting different types of information into our textual prompts. In *Word from Q*, the marker contains the first word from the historic answer's corresponding question, which is typically a wh-word (e.g., ''<what>''). In *Word from Q + Index* we also add the historic answer's turn index (e.g., ''<what_1>''). In *Full Q*, we inject the entire historic question into the prompt. *Word from Q* and *Word from Q + Index* achieved comparable scores, lower than MarCQAp's but higher than *Answer Pos*'s.[21] This suggests that adding semantic information is useful (since *Word from Q* outperformed *Answer Pos*), and that combining such information with the positional information is not trivial (since MarCQAp outperformed *Word from Q + Index*). This points to the effects of the prompt structure and the information included: We see that ''<1>'' and ''<what>'' both outperform ''<>'', yet constructing a prompt by naively combining these signals (''<what_1>'') does not lead to complementary effect. Finally, *Word from Q* outperformed *Full Q*. We hypothesize that since the full question can be long, it might substantially interfere with the natural structure of the passage text. This provides evidence that the prompts should probably remain compact symbols with small footprint within the passage. These initial results call for further exploration of optimal prompt design in future work.

5.4 Case Study

Figure 5 presents an example of all evaluated methods in action from the *standard* setting. The current question *''Did he have any other **critics**?''* has two correct answers: ***Alan Dershowitz*** or ***Omer Bartov***. We first note that all methods predicted a name of a person, which indicates that the main subject of the question was captured correctly. Yet, the methods differ in their prediction of the specific person.



QuAC id: C_b728d731c83e4376959ce4db09fed0b7_0.

Wikipedia Title: Norman Finkelstein.

Passage:
*Criticism has been leveled against Finkelstein **… … …** Daniel Goldhagen, **…** , claimed his scholarship has "everything to do with his burning political agenda". Alan Dershowitz has written that Peter Novick, Professor of History **…** whose work Finkelstein says inspired The Holocaust Industry, has strongly criticized the latter's work, describing it as "trash". Similarly, Dershowitz, whose book **…** , has claimed Finkelstein complicity in a conspiracy **… … …** Israeli historian Omer Bartov, … , judged The Holocaust Industry to be marred by the same errors … : It is filled with precisely the kind of shrill hyperbole that Finkelstein rightly deplores in much of the current media hype over the Holocaust; **… …***

Q1: What were Norman's criticisms?
A1: *Daniel Goldhagen, **…** , claimed his scholarship has "everything to do with his burning political agenda".*
Q2: Which other critics does he have?
A2: *Peter Novick, Professor of History at **…***
Q3: How does he criticize him?
A3: *strongly criticized the latter's work, describing it as "trash".*
Q4: Did he have any other critics?

Rewritten: Besides Peter Novick, did Norman Finkelstein have any other critics?
Gold Rewritten: Did Norman Finkelstein have any other critics aside from Peter Novick and Daniel Goldhagen?

Correct Answers:
- *Israeli historian Omer Bartov,*
- *Dershowitz, whose book …*

Concat	*Alan Dershowitz has written that Peter Novick, …*
Rewrite	*Dershowitz, whose book …*
RewriteC	*Daniel Goldhagen*
ExCorD	*Finkelstein has accused journalist Jeffrey Goldberg …*
HAE	*Finkelstein has accused journalist Jeffrey Goldberg …*
PosHAE	*Finkelstein has accused journalist Jeffrey Goldberg …*
MarCQAp	*Israeli historian Omer Bartov,*

Figure 5: Our case study example, comparing answers predicted by each evaluated method in the *standard* setting. We provide a detailed analysis in §5.4.

Rewrite and Concat predict a correct answer (*Alan Dershowitz*), yet Concat predicts it based on incorrect evidence. This may indicate that Concat did not capture the context correctly (just the fact that it needs to predict a person's name), and was lucky enough to guess the correct name.

Interestingly, Rewrite$_C$ predicts *Daniel Goldhagen*, which is different from the answers predicted by Concat and Rewrite. This shows that combining both methods can yield completely different results, and demonstrates an instance where Rewrite$_C$ performs worse than Rewrite and Concat (for instance in the 1% *low-resource* setting). This is also an example of a history modeling flaw, since *Daniel Goldhagen* was already mentioned as a critic in previous conversation turns.

This example also demonstrates how errors can propagate through a pipeline-based system.

[21] Both differences are statistically significant.

The gold rewritten question is *''Did Norman Finkelstein have any other critics aside from **Peter Novick** and **Daniel Goldhagen**?''*,[22] while the question rewriting model generated *''Besides **Peter Novick**, did Norman Finkelstein have any other critics?''*, omitting **Daniel Goldhagen**. This makes it impossible for REWRITE to figure out that **Daniel Goldhagen** was already mentioned, making it a legitimate answer. This reveals that REWRITE might have also gotten lucky and provides a possible explanation for the incorrect answer predicted by REWRITE$_C$.

ExCorD$_{LF}$, HAE$_{LF}$, and PosHAE$_{LF}$ not only predict a wrong answer, but also seem to fail to resolve the conversational coreferences, since the pronoun *''**he**''*, in the current question *''Did **he** have any other critics?''*, refers to ***Norman Finkelstein***.

MarCQAp predicts a correct answer, *Omer Bartov*. This demonstrates an instance where MarCQAp succeeds while HAE$_{LF}$ and PosHAE$_{LF}$ fail, even though they are all history-highlighting methods. Interestingly, MarCQAp is the only model that predicts *Omer Bartov*, a non-trivial choice compared to *Alan Dershowitz*, since *Omer Bartov* appears later in the passage, further away from the historic answers.

6 Limitations

This work focuses on a single-document CQA setting, which is in line with the majority of the previous work on conversation history modeling in CQA (§2.3). Correspondingly, MarCQAp was designed for single-document CQA. Applying MarCQAp in multi-document settings (Qu et al., 2020; Anantha et al., 2021; Adlakha et al., 2022) may result in partial history representation, since the retrieved document may contain only part of the historic answers, therefore MarCQAp will only highlight the answers which appear in the document.[23]

In §5.3 we showed initial evidence that MarCQAp prompts can encode additional information that can be useful for CQA. In this work we focused on the core idea behind prompt-based answer highlighting, as a proposed solution in light of our results in §4. Yet, we did not conduct a comprehensive exploration in search of the optimal prompt design, and leave this for future work.

7 Conclusion

In this work, we carry out the first comprehensive robustness study of history modeling approaches for Conversational Question Answering (CQA), including sensitivity to model and training data size, domain shift, and noisy history input. We revealed limitations of the existing benchmark-based evaluation, by demonstrating that it cannot reflect the models' robustness to such changes in setting. In addition, we proposed evaluation practices that better isolate the contribution of the history modeling component, and demonstrated their usefulness.

We also discovered that highlighting historic answers via passage embedding is very effective in *standard* setups, but it suffers from substantial performance degradation in low data and domain shift settings. Following this finding, we design a novel prompt-based history highlighting approach. We show that highlighting with prompts, rather than with embeddings, significantly improve robustness, while maintaining overall high performance.

Our approach can be a good starting point for future work, due to its high effectiveness, robustness, and portability. We also hope that the insights from our study will encourage evaluations with focus on robustness, leading to better CQA systems.

Acknowledgments

We would like to thank the action editor and the reviewers, as well as the members of the IE@Technion NLP group and Roee Aharoni for their valuable feedback and advice. The Technion team was supported by the Zuckerman Fund to the Technion Artificial Intelligence Hub (Tech.AI). This research was also supported in part by a grant from Google.

References

Vaibhav Adlakha, Shehzaad Dhuliawala, Kaheer Suleman, Harm de Vries, and Siva Reddy. 2022. Topiocqa: Open-domain conversational question answering with topic switching. *Transactions of the Association for Computational Linguistics*, 10:468–483. https://doi.org/10.1162/tacl_a_00471

[22] As annotated in CANARD (Elgohary et al., 2019).

[23] We note that this limitation applies to all highlighting approaches, including HAE and PosHAE (Qu et al., 2019a,b).

Armen Aghajanyan, Dmytro Okhonko, Mike Lewis, Mandar Joshi, Hu Xu, Gargi Ghosh, and Luke Zettlemoyer. 2022. HTLM: Hypertext pre-training and prompting of language models. In *International Conference on Learning Representations*.

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, NAACL-HLT 2021, Online, June 6–11, 2021*, pages 520–534. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.44

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *CoRR*, abs/2004.05150.

Eyal Ben-David, Nadav Oved, and Roi Reichart. 2022. PADA: Example-based prompt learning for on-the-fly adaptation to unseen domains. *Transactions of the Association for Computational Linguistics*, 10:414–433. https://doi.org/10.1162/tacl_a_00468

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jon Ander Campos, Arantxa Otegi, Aitor Soroa, Jan Deriu, Mark Cieliebak, and Eneko Agirre. 2020. Doqa - accessing domain-specific faqs via conversational QA. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5–10, 2020*, pages 7302–7314. Association for Computational Linguistics.

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. https://doi.org/10.24963/ijcai.2020/171

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau 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. https://doi.org/10.18653/v1/D18-1241

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15–20, 2018, Volume 1: Long Papers*, pages 1383–1392. Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1128

Rotem Dror, Lotem Peled-Cohen, Segev Shlomov, and Roi Reichart. 2020. *Statistical Significance Testing for Natural Language Processing*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers. https://doi.org/10.1007/978-3-031-02174-9

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. https://doi.org/10.18653/v1/D19-1605

Somil Gupta, Bhanu Pratap Singh Rawat, and Hong Yu. 2020. Conversational machine comprehension: A literature review. In

Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8–13, 2020, pages 2739–2753. International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.247

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.

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. https://doi.org/10.18653/v1/2021.acl-long.478

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*.

Teven Le Scao and Alexander Rush. 2021. How many data points is a prompt worth? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2627–2636, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.208

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. https://doi.org/10.18653/v1/2022.acl-long.555

Sheng-Chieh Lin, Jheng-Hong Yang, Rodrigo Nogueira, Ming-Feng Tsai, Chuan-Ju Wang, and Jimmy Lin. 2020. Conversational question reformulation via sequence-to-sequence architectures and pretrained language models. *CoRR*, abs/2004.01909.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *CoRR*, abs/2107.13586.

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. *CoRR*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6–9, 2019*. OpenReview.net.

Jin Qian, Bowei Zou, Mengxing Dong, Xiao Li, AiTi Aw, and Yu Hong. 2022. Capturing conversational interaction for question answering via global history reasoning. In *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10–15, 2022*, pages 2071–2078. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-naacl.159

Chen Qu, Liu Yang, Cen-Chieh Chen, Minghui Qiu, W. Bruce Croft, and Mohit Iyyer. 2020. Open-retrieval conversational question answering. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:140:1–140:67.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15–20, 2018, Volume 2: Short Papers*, pages 784–789. Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-2124

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. https://doi.org/10.18653/v1/D16-1264

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266. https://doi.org/10.1162/tacl_a_00266

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4–9, 2017, Long Beach, CA, USA*, pages 5998–6008.

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. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Yi-Ting Yeh and Yun-Nung Chen. 2019. Flowdelta: Modeling flow information gain in reasoning for conversational machine comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering, MRQA@EMNLP 2019, Hong Kong, China, November 4, 2019*, pages 86–90. Association for Computational Linguistics.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6–12, 2020, virtual*.

Jing Zhao, Junwei Bao, Yifan Wang, Yongwei Zhou, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2021. Ror: Read-over-read for long document machine reading comprehension. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16–20 November, 2021*, pages 1862–1872. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-emnlp.160

Chenguang Zhu, Michael Zeng, and Xuedong Huang. 2018. Sdnet: Contextualized attention-based deep network for conversational question answering. *CoRR*, abs/1812.03593.