Post-Abstention: Towards Reliably Re-Attempting the Abstained Instances in QA

Neeraj Varshney and Chitta Baral
Arizona State University

Abstract

Despite remarkable progress made in natural language processing, even the state-of-the-art models often make incorrect predictions. Such predictions hamper the reliability of systems and limit their widespread adoption in real-world applications. *Selective prediction* partly addresses the above concern by enabling models to abstain from answering when their predictions are likely to be incorrect. While selective prediction is advantageous, it leaves us with a pertinent question '*what to do after abstention*'. To this end, we present an explorative study on 'Post-Abstention', a task that allows re-attempting the abstained instances with the aim of increasing *coverage* of the system without significantly sacrificing its *accuracy*. We first provide mathematical formulation of this task and then explore several methods to solve it. Comprehensive experiments on 11 QA datasets show that these methods lead to considerable risk improvements –performance metric of the Post-Abstention task– both in the in-domain and the out-of-domain settings. We also conduct a thorough analysis of these results which further leads to several interesting findings. Finally, we believe that our work will encourage and facilitate further research in this important area of addressing the reliability of NLP systems.

1 Introduction

Despite remarkable progress made in Natural Language Processing (NLP), even the state-of-the-art systems often make incorrect predictions. This problem becomes worse when the inputs tend to diverge from the training data distribution (Elsahar and Gallé, 2019; Miller et al., 2020; Koh et al., 2021). Incorrect predictions hamper the reliability of systems and limit their widespread adoption in real-world applications.

Selective prediction partly addresses the above concern by enabling models to abstain from answering when their predictions are likely to be incorrect. By avoiding potentially incorrect predictions, it allows maintaining high task accuracy and thus improves the system's reliability. Selective prediction has recently received considerable attention from the NLP community leading to development of several methods (Kamath et al., 2020; Garg and Moschitti, 2021; Xin et al., 2021; Varshney et al., 2022d). While these contributions are important, selective prediction leaves us with a pertinent question: *what to do after abstention?*

In this work, we address the above question and present an explorative study on '**Post-Abstention**', a task that allows re-attempting the abstained instances with the aim of increasing *coverage* of the given selective prediction system without significantly sacrificing its *accuracy*. Figure 1 illustrates the benefit of employing a post-abstention method; a model that achieves an accuracy of 70% is first enabled with the selective prediction ability that increases the accuracy to 85% but answers only 71% instances. Then, a post-abstention method is employed (for the 29% abstained instances) that assists the system in answering 9% more instances raising the coverage to 80% without considerably dropping the overall accuracy. We note that this task allows re-attempting all the abstained instances but does not require the system to necessarily output predictions for all of them i.e. the system can abstain even after utilizing a post-abstention method (when it is not sufficiently confident even in its new prediction). This facet not only allows the system to maintain its performance but also provides opportunities of sequentially applying stronger post-abstention methods to reliably and optimally increase the coverage in stages.

We provide mathematical formulation of the post-abstention task and explore several baseline methods to solve it (Section 2). To evaluate the efficacy of these methods, we conduct comprehensive experiments with 11 Question-Answering datasets from MRQA shared task (Fisch et al., 2019) in

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 967–982

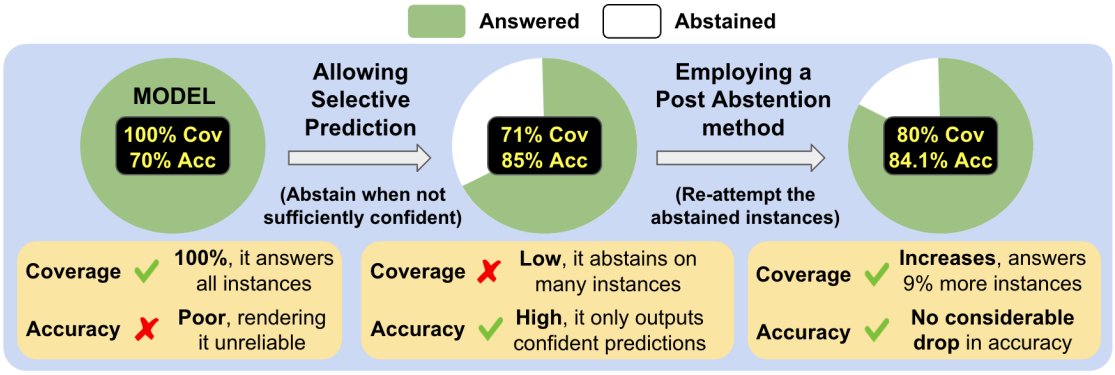

Figure 1: Illustrating the **impact of employing a post-abstention method** on top of selective prediction system. A regular model that has an accuracy of 70% (at coverage 100%) is first enabled with selective prediction ability that increases the accuracy to 85% but drops the coverage to 71%. Then, on employing a post-abstention method to the abstained instances (remaining 29%), coverage increases to 80% without a considerable drop in overall accuracy.

both in-domain and out-of-domain settings (Section 3). Our post-abstention methods lead to overall risk improvements (performance metric of the proposed task) of up to 21.81 in the in-domain setting and 24.23 in the out-of-domain setting. To further analyze these results, we study several research questions, such as 'what is the extent of overlap between the instances answered by different post-abstention methods', 'what is the distribution of model's original confidence on instances that get answered in the post-abstention stage', and 'how often do the system's predictions change after applying post-abstention methods'. In Section 4, we show that these investigations lead to numerous important and interesting findings.

In summary, our contributions are as follows:

1. We present an **explorative study on 'Post-Abstention'**, a task that aims at increasing the *coverage* of a given selective prediction system without significantly sacrificing its *accuracy*.
2. We **explore several baseline post-abstention methods** and evaluate them in an extensive experimental setup spanning 11 QA datasets in both in-domain and out-of-domain settings.
3. We show that the proposed post-abstention methods **result in overall risk value improvements** of up to 21.81 and 24.23 in the in-domain and out-of-domain settings respectively.
4. Our **thorough analysis** leads to several interesting findings, such as (a) instances answered by different post-abstention methods are not mutually exclusive i.e. there exist some overlapping instances, (b) instances that get answered in the post-abstention stage are not necessarily the ones on which the given system was initially most confident, etc.

We believe our work will encourage further research in Post-Abstention, an important step towards improving the reliability of NLP systems.

2 Post-Abstention

In this section, we first provide background for post-abstention (2.1) and then describe the task (2.2) and its approaches (2.3).

2.1 Background

Post-abstention, as the name suggests, is applicable for a system that abstains from answering i.e. a selective prediction system. A system can typically abstain when its prediction is likely to be incorrect. This improves the reliability of the system. Such a system typically consists of two functions: a predictor (f) that gives the model's prediction on an input (x) and a selector (g) that determines if the system should output the prediction made by f:

$$(f, g)(x) = \begin{cases} f(x), & \text{if g(x) = 1} \\ Abstain, & \text{if g(x) = 0} \end{cases}$$

Typically, g comprises of a prediction confidence estimator $\tilde{g}$ and a threshold th that controls the level of abstention for the system:

$$g(x) = \mathbb{1}[\tilde{g}(x)) > th]$$

A selective prediction system makes trade-offs between $coverage$ and $risk$. Coverage at a threshold th is defined as the fraction of total instances answered by the system (where $\tilde{g} > th$) and risk is the error on the answered instances.

With decrease in threshold, coverage will increase, but the risk will usually also increase. The

overall selective prediction performance is measured by the *area under Risk-Coverage curve* (El-Yaniv et al., 2010) which plots risk against coverage for all confidence thresholds. Lower AUC is better as it represents lower average risk across all confidence thresholds.

In NLP, approaches such as Monte-Carlo Dropout (Gal and Ghahramani, 2016), Calibration (Kamath et al., 2020; Varshney et al., 2022c,d; Zhang et al., 2021), Error Regularization (Xin et al., 2021) and Label Smoothing (Szegedy et al., 2016) have been studied for selective prediction. In this work, we consider MaxProb (Hendrycks and Gimpel, 2017), a technique that uses the maximum softmax probability across all answer candidates as the confidence estimator. We use this simple technique because the focus of this work is on post-abstention i.e. the next step of selective prediction. However, we note that the task formulation and the proposed methods are general and applicable to all selective prediction approaches.

2.2 Task Formulation

We define the post-abstention task as follows:

> *Given a selective prediction system with an abstention threshold, the post-abstention task allows re-attempting the abstained instances with the aim of improving the coverage without considerably degrading the accuracy (or increasing the risk)* of the given system.

Next, we mathematically describe the task and its performance evaluation methodology.

Let the coverage and risk of the given selective prediction system at abstention threshold th be cov_{th} and $risk_{th}$ respectively. A post-abstention method re-attempts the originally abstained instances (where $\tilde{g} < th$) and outputs the new prediction for the ones where it is now sufficiently confident. This typically leads to an increase in the coverage of the system with some change in the risk value; let the new coverage and risk be cov'_{th} and $risk'_{th}$ respectively. From the risk-coverage curve of the given system, we calculate its risk at coverage cov'_{th} and compare it with $risk'_{th}$ to measure the efficacy of the post-abstention method (refer to Figure 2).

For a method to have a positive impact, its risk ($risk'_{th}$) should be lower than the risk of the given system at coverage cov'_{th}. We summarize this performance evaluation methodology in Figure 2. To get an overall performance estimate of a post-abstention method, we compile these differences in risk values for all confidence thresholds and calculate an aggregated value. The higher the overall improvement value, the more effective the method is. We note that this evaluation methodology is fair and accurate as it conducts pair-wise comparisons at **equal coverage** points. An alternative performance metric could be AUC but it computes the overall area ignoring the pair-wise comparisons which are crucial for our task because the coverage points of the original system would be different from those achieved by the post-abstention method.

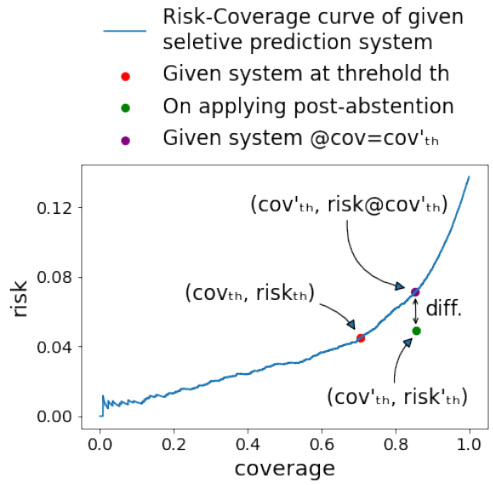

Figure 2: Summarizing **performance evaluation methodology** of post-abstention. Given a selective prediction system with coverage cov_{th} and risk $risk_{th}$ at abstention threshold th, let the new coverage and risk after applying a post-abstention method be cov'_{th} and $risk'_{th}$ respectively. From the risk-coverage curve of the given system, we calculate its risk at coverage cov'_{th} and compare it with $risk'_{th}$ (diff). *For the method to have a positive impact, $risk'_{th}$ should be lower than the risk of the given system at coverage cov'_{th}.*

2.3 Approaches

2.3.1 Ensembling using Question Paraphrases

It is well known that even state-of-the-art NLP models are often brittle i.e. when small semantic-preserving changes are made to the input, their predictions tend to fluctuate greatly (Jia and Liang, 2017; Belinkov and Bisk, 2018; Iyyer et al., 2018; Ribeiro et al., 2018; Wallace et al., 2019). Ensembling the predictions of the model on multiple semantically equivalent variants of the input is a promising approach to address this issue (Anantha et al., 2021; Vakulenko et al., 2021) as it can reduce the spread or dispersion of the predictions.

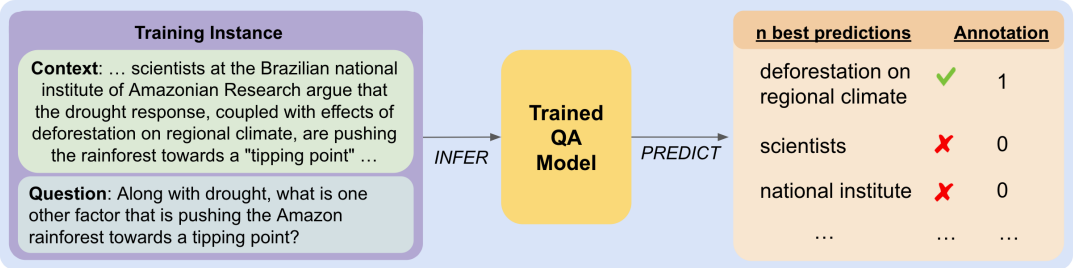

Figure 3: Illustrating **annotation procedure of REToP**. For each training instance, top N predictions given by the QA model are annotated conditioned on their correctness i.e. correct predictions are annotated as '1' and incorrect predictions are annotated as '0'. This annotated binary classification dataset is used to train the auxiliary model.

We leverage the above technique in re-attempting the abstained questions i.e. we first generate multiple paraphrases of the input instance and then aggregate the model's predictions on them. We use BART-large (Lewis et al., 2019) model fine-tuned on Quora Question Corpus (Iyer et al., 2017), PAWS (Zhang et al., 2019), and Microsoft Research Paraphrase Corpus (Dolan and Brockett, 2005) for paraphrasing and explore the following strategies for aggregating the model predictions:

- **Mean**: In this strategy, we calculate the average confidence assigned to each answer candidate across all predictions. Then, we select the candidate with the highest average confidence as the system's prediction. Note that the system will output this prediction only if its confidence surpasses the abstention threshold.

- **Max**: Here, like the *mean* strategy, we select the answer candidate with the highest average confidence but we use the maximum confidence assigned to that candidate as its prediction confidence. This is done to push the most confident prediction above the abstention threshold.

2.3.2 Re-Examining Top N Predictions (REToP)

State-of-the-art models have achieved impressive performance on numerous NLP tasks. Even in cases where they fail to make a correct prediction, they are often able to rank the correct answer as one of their top N predictions. This provides opportunities for re-examining the top N predictions to identify the correct answer in case of abstention. To this end, a model that can estimate the correctness of a prediction can be leveraged. Following this intuition, we develop an **auxiliary model** that takes the context, question, and a prediction as input and assigns a score indicating the likelihood of that prediction to be correct. This model can be used for each of the top N predictions given by the QA model to select the one that is most likely to be the correct answer.

Training Auxiliary Model: We first create data instances by annotating (context, question, prediction) triplets conditioned on the correctness of the QA system's predictions and then train a classification model using this data. This model is specific to the given QA system and essentially learns to distinguish its correct and incorrect predictions.

- **Annotate (context, question, prediction) triplets**: We utilize the trained QA model to get its top N predictions for each training instance. Then, we annotate each (context, question, prediction) triplet based on the prediction's correctness i.e. a correct prediction is annotated as '1' and an incorrect prediction is annotated as '0'. Figure 3 illustrates this annotation step.

- **Train a classification model**: Then, a binary classification model is trained using the annotated dataset collected in the previous step. This model specifically learns to distinguish the correct predictions of the QA model from the incorrect ones. Softmax probability assigned to the label '1' corresponds to the likelihood of correctness for each prediction.

Note that we use the QA model's top N predictions to collect the '0' annotations instead of randomly selecting candidates because this procedure results in highly informative negative instances (that are probable predictions and yet incorrect) and not easy/obvious negatives. This can help the auxiliary model in learning fine-grained representations distinguishing correct and incorrect predictions.

Leveraging Auxiliary Model: For an abstained instance, we compute the likelihood value for each

of the top N predictions given by the QA model using our trained auxiliary model. Then, we calculate the overall confidence (c) of each prediction (p) as a weighted average of the QA model's probability (s_q) and the auxiliary model's likelihood score (s_a) i.e. c_p is calculated as:

$$c_p = \alpha * s_q^p + (1 - \alpha) * s_a^p$$

where α is a weight parameter.

We incorporate QA model's probability as it provides more flexibility to compute the overall confidence. Finally, prediction with the highest overall confidence is selected as the new prediction. We differentiate this method from existing methods such as calibration in Appendix C.

2.3.3 Human Intervention (HI)

In intolerant application domains such as biomedicals where incorrect predictions can have serious consequences, human intervention is the most reliable technique to answer the abstained instances. Human intervention can be in various forms such as providing relevant knowledge to the model, asking clarifying questions (Rao and Daumé III, 2018) or simplifying the input question. In this work, we explore a simple human intervention approach in which the system provides multiple predictions instead of only one prediction for the abstained instances. The human can then select the most suitable prediction from the provided predictions. Performance of this method can be approximated based on the presence of the correct answer in the predictions provided to the human. Note that the above approach would answer all the abstained instances and hence the coverage would always be 100%. This implies that with the increase in abstention threshold, the risk would monotonically decrease as multiple predictions would be returned for a larger number of instances.

In addition to the above approach, we also explore a **REToP-centric** HI approach in which the system returns multiple predictions only when REToP surpasses the confidence threshold in the post-abstention stage. Similar to REToP, it abstains on the remaining instances. Finally, we note that comparing the performance of HI approaches with other post-abstention approaches would be unfair as other approaches return only a single prediction. Therefore, we present HI results separately.

3 Experiments and Results

3.1 Experimental Setup

Datasets: We experiment with SQuAD 1.1 (Rajpurkar et al., 2016) as the source dataset and the following 10 datasets as out-of-domain datasets: NewsQA (Trischler et al., 2017), TriviaQA (Joshi et al., 2017), SearchQA (Dunn et al., 2017), HotpotQA (Yang et al., 2018), and Natural Questions (Kwiatkowski et al., 2019), DROP (Dua et al., 2019), DuoRC (Saha et al., 2018), RACE (Lai et al., 2017), RelationExtraction (Levy et al., 2017), and TextbookQA (Kim et al., 2019). We use the preprocessed data from the MRQA shared task (Fisch et al., 2019) for our experiments.

Implementation Details: We run all our experiments using the huggingface (Wolf et al., 2020) implementation of transformers on Nvidia V100 16GB GPUs with a batch size of 32 and learning rate ranging in $\{1-5\}e-5$. We generate 10 paraphrases of the question in Ensembling method, re-examine top 10 predictions, vary α in the range $0.3 - 0.7$ for REToP method, and vary the number of predictions in the range 2 to 5 for HI methods. Since the focus of this work is on post-abstention, it's crucial to experiment with models that leave sufficient room for effectively evaluating the ability of post-abstention methods. For that reason, we experiment with a small size model (BERT-mini having just 11.3M parameters) from Turc et al. (2019) for our experiments. However, we note that our methods are general and applicable for all models.

3.2 Results

3.2.1 REToP

Table 1 shows the post-abstention performance of REToP for selected abstention thresholds. The last column ('*Total Risk Improvement*') in this table corresponds to the overall improvement aggregated over all confidence thresholds. It can be observed that REToP achieves considerable risk improvements both in the in-domain setting (21.81 on SQuAD) and the out-of-domain settings (24.23 on TextbookQA, 21.54 on HotpotQA, 20.42 on RE, etc). Next, we analyze these results in detail.

Higher improvement on moderate confidences: In Figure 4, we plot risk improvements achieved by REToP on SQuAD (in-domain) and HotpotQA (out-of-domain) datasets for all confidence thresholds. These plots reveal that the improvement is

Dataset	Model	0.2		0.32		0.36		0.48		0.54		0.60		0.68		Total Risk
		Cov↑	Risk↓	Cov↑	Risk↓	Cov↑	Risk↓	Cov↑	Risk↓	Cov↑	Risk↓	Cov↑	Risk↓	Cov↑	Risk↓	Improvement↑
	Given (G)	96.65	32.45	87.24	28.10	83.34	26.69	69.94	21.91	62.57	19.91	56.23	17.98	47.92	15.43	
SQuAD	RETOP	99.73	**33.75**	97.27	**31.93**	95.08	**30.85**	80.88	**24.84**	72.44	**21.82**	63.73	**19.19**	52.65	**16.43**	
(in-domain)	G@RETOP$_{cov}$	-	34.00	-	32.77	-	31.67	-	25.82	-	22.59	-	20.24	-	16.83	**21.81**
	Given (G)	97.54	67.65	89.56	65.88	85.39	65.13	71.75	62.71	64.77	61.56	58.19	60.34	49.25	58.29	
HotpotQA	RETOP	99.93	**68.17**	98.63	**67.39**	96.9	**66.61**	82.88	**63.61**	73.55	**61.89**	64.36	**60.53**	52.96	**58.34**	
	G@RETOP$_{cov}$	-	68.30	-	67.92	-	67.47	-	64.52	-	63.04	-	61.55	-	59.01	**21.54**
	Given (G)	97.59	44.49	89.01	40.51	85.41	39.04	74.08	34.16	66.86	30.54	60.58	27.94	54.10	24.20	
RE	RETOP	99.93	**45.38**	98.95	**44.39**	97.52	**43.79**	85.89	**38.67**	77.61	**34.57**	69.54	**31.12**	59.33	**25.39**	
	G@RETOP$_{cov}$	-	45.47	-	45.01	-	44.43	-	39.22	-	35.51	-	32.10	-	27.33	**20.42**
	Given (G)	89.02	80.5	71.07	77.04	66.17	75.56	51.34	72.54	43.47	69.62	36.2	68.85	29.97	63.86	
RACE	RETOP	99.41	82.24	92.28	**80.71**	86.94	**79.35**	62.91	**73.82**	51.48	**71.76**	42.28	**69.47**	33.09	**65.92**	
	G@RETOP$_{cov}$	-	81.94	-	81.00	-	80.00	-	75.00	-	72.54	-	69.72	-	66.37	**15.10**
	Given (G)	93.90	69.76	80.91	66.40	75.5	64.91	60.30	60.79	53.30	58.8	47.17	56.62	39.32	54.11	
NewsQA	RETOP	99.48	**71.03**	96.13	**70.24**	93.21	69.64	70.85	**63.71**	60.73	**60.67**	52.04	**58.07**	42.09	**54.94**	
	G@RETOP$_{cov}$	-	71.31	-	70.36	-	69.61	-	63.81	-	61.01	-	58.33	-	55.02	**5.10**
	Given (G)	96.15	86.68	81.77	85.67	75.77	85.34	58.64	84.08	50.22	83.58	42.67	83.33	34.46	82.55	
SearchQA	RETOP	99.92	87.06	97.58	86.81	93.92	**86.48**	71.49	**84.76**	59.46	**84.04**	48.6	**83.48**	37.08	**82.75**	
	G@RETOP$_{cov}$	-	87.04	-	86.79	-	86.52	-	85.07	-	84.15	-	83.56	-	82.77	**1.78**
	Given (G)	96.67	67.31	86.89	65.05	82.54	63.82	68.81	60.39	61.44	58.39	55.11	56.48	47.12	54.03	
TriviaQA	RETOP	99.86	**68.07**	97.07	**67.33**	93.72	**66.23**	76.72	62.40	67.93	60.25	59.55	**57.77**	49.29	54.89	
	G@RETOP$_{cov}$	-	68.09	-	67.42	-	66.60	-	62.32	-	60.12	-	57.95	-	54.83	**0.70**
	Given (G)	92.37	63.78	79.04	59.99	74.87	58.77	60.60	53.51	54.03	51.00	47.94	48.31	41.70	45.27	
NQ	RETOP	98.71	**65.34**	93.04	**63.39**	89.30	**62.62**	70.65	**56.90**	61.68	**53.54**	53.24	**50.10**	43.75	**46.44**	
	G@RETOP$_{cov}$	-	65.67	-	63.93	-	63.02	-	57.43	-	53.80	-	50.68	-	46.45	**10.70**
	Given (G)	95.74	88.46	81.17	87.38	76.11	87.33	62.34	86.23	53.69	85.38	48.77	84.45	43.05	85.01	
DROP	RETOP	99.53	88.64	92.95	**87.83**	88.42	88.04	69.00	**86.31**	58.55	**85.57**	51.90	**84.49**	44.18	85.09	
	G@RETOP$_{cov}$	-	88.63	-	88.19	-	87.88	-	86.69	-	85.91	-	84.87	-	84.94	**3.63**
	Given (G)	97.20	68.68	87.87	66.41	84.21	65.82	71.09	62.42	64.16	61.47	57.16	59.91	50.03	58.46	
DuoRC	RETOP	99.87	**69.45**	98.33	**69.17**	96.14	68.68	80.75	**64.69**	71.95	**62.59**	62.56	**60.70**	52.90	**58.69**	
	Original@cov	-	69.51	-	69.02	-	68.4	-	64.77	-	62.74	-	60.92	-	59.32	**4.32**
	Given (G)	94.34	67.14	80.9	63.32	75.65	61.92	57.49	56.02	49.63	52.14	41.45	51.04	34.07	50.00	
TBQA	RETOP	99.53	**68.38**	95.01	**67.23**	91.68	**66.18**	68.20	**58.34**	58.55	**54.77**	47.37	**51.26**	37.26	**49.64**	
	G@RETOP$_{cov}$	-	68.56	-	67.30	-	66.23	-	59.41	-	56.02	-	52.60	-	50.71	**24.23**

Table 1: **Performance of RETOP as a post-abstention method** for selected abstention thresholds. The QA model is trained using SQuAD training data and evaluated on SQuAD (in-domain) and 10 out-of-domain datasets. For each dataset, we provide three rows: first row ('*Given*') shows the coverage and risk values of the given selective prediction system at different abstention thresholds, second row ('*RETOP*') shows the coverage and risk after applying RETOP on abstained instances of the given system, and third row ('*G@RETOP$_{cov}$*') shows risk of the given system at the coverage achieved by *RETOP*. For the post abstention method to be effective, risk in the second row should be less than that in the third row and the magnitude of difference corresponds to the improvement. The last column corresponds to the overall improvement aggregated over all confidences ranging from 0 to 1 at an interval of 0.02. ↓ and ↑ indicate that lower (risk) and higher (coverage, risk improvement) values are better respectively.

more on moderate thresholds as compared to low thresholds. We attribute this to the high difficulty of instances that remain to be re-attempted at low thresholds i.e. only the instances on which the given system was highly underconfident are left for the post-abstention method. It has been shown that model's confidence is negatively correlated with difficulty (Swayamdipta et al., 2020; Rodriguez et al., 2021; Varshney et al., 2022b) implying that the remaining instances are tough to be answered correctly. This justifies the lesser improvement in performance observed at low thresholds.

In-Domain vs Out-of-Domain Improvement: RETOP achieves higher performance improvement on the in-domain dataset than the out-of-domain datasets (on average). This is expected as the auxiliary model in RETOP is trained using the in-domain training data. However, it still has good performance on out-of-domain datasets as the auxiliary model learns fine-grained representations to distinguish between correct and incorrect predictions. Furthermore, the improvement on out-of-domain

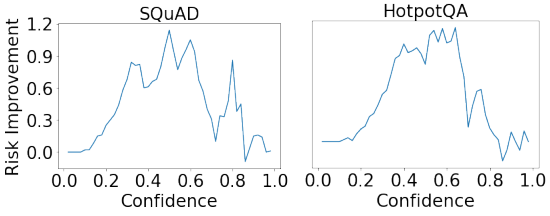

Figure 4: **Improvement in risk** achieved by using RETOP in post-abstention on SQuAD (in-domain) and HotpotQA (out-of-domain) datasets for all confidences.

Dataset	Ens.	RETOP ($\alpha = 0.6$)	RETOP ($\alpha = 0.65$)	*HI on (RETOP)
SQuAD	0.29	21.81	20.02	47.85
HotpotQA	0.93	21.54	19.00	37.88
RE	21.72	20.42	17.61	46.65
RACE	16.72	15.10	14.17	36.26
NewsQA	11.92	5.10	5.10	26.41
SearchQA	17.05	1.78	2.23	20.08
TriviaQA	9.50	0.70	1.47	17.21
NQ	13.40	10.70	10.89	31.95
DROP	1.57	3.63	2.99	8.08
DuoRC	-1.69	4.32	5.90	20.26
TBQA	-6.93	24.23	23.73	45.18
Total	84.48	**129.33**	123.11	337.81

Table 2: **Comparing total risk improvement** achieved by different post-abstention methods. * for HI indicates that it's results are not directly comparable as it outputs multiple predictions while others output only one.

Dataset	0.0	0.2	0.4	0.6	0.8
SQuAD	34.15	33.72	30.9	28.05	26.3
HotpotQA	68.33	68.19	66.56	63.65	61.57
RE	45.52	45.35	43.39	41.28	39.31
RACE	82.05	81.6	80.12	78.19	77.15
NewsQA	71.46	71.2	69.42	67.21	65.29
SearchQA	87.06	86.92	85.64	83.98	82.94
TriviaQA	68.13	67.9	66.62	64.21	62.47
NQ	66.09	65.67	63.63	61.06	59.31
DROP	88.69	88.69	87.56	86.36	85.7
DuoRC	69.55	69.42	68.15	66.42	65.22
TBQA	68.73	68.46	67.07	64.74	64.01

Table 3: Comparing **risk values achieved by the HI method** (returns two predictions for all abstained instances) across different abstention thresholds.

data varies greatly across datasets (from 0.7 on TriviaQA to 24.23 on TextbookQA).

3.2.2 Comparing Post-Abstention Approaches

We provide the performance tables for other post-abstention approaches in Appendix. However, we compare their total risk improvement values in Table 2. In the in-domain setting, RETOP achieves higher improvement than Ensembling method. This is because the auxiliary model in RETOP has specifically learned to distinguish the correct and incorrect predictions from the training data of this domain. However, in some out-of-domain cases, Ensembling outperforms RETOP (SearchQA, TriviaQA, NewsQA). Overall, RETOP leads to a consistent and higher risk improvement on average. Ensembling also leads to a minor degradation in a few out-of-domain datasets (DuoRC and TextbookQA). Next, we analyze the performance of human intervention (HI) methods.

3.2.3 Human Intervention (HI)

We study two variants of HI method. In the first variant, multiple predictions (n=2) are returned for all the abstained instances. This makes the coverage to be 100% for all the confidences; therefore, we present only the risk values in Table 3. As expected, with increase in abstention threshold, the risk decreases because multiple predictions get outputted for a larger number of instances. Selection of operating threshold for an application depends on the trade-off between risk that can be tolerated and human effort required to select the most suitable prediction from a set of predictions returned by the system. For example, a low threshold can be selected for tolerant applications like movie recommendations and a high threshold for tolerant applications like house robots.

In the second variant of HI method, we study a **RETOP-centric** approach in which the system returns multiple predictions only when RETOP surpasses the confidence threshold in the post-abstention stage. The last column in Table 2 shows the risk improvements achieved by this approach (n=2). Note that RETOP re-examines the top N predictions and selects one while this method outputs multiple predictions and requires a human to select the most suitable one. These results indicate that though RETOP achieves good performance, there is still some room for improvement.

3.2.4 Ensembling Using Paraphrases

Comparing the performance of Mean and Max Ensembling strategies reveals that Max increases the coverage more than the Mean strategy but it also increases the risk considerably. Thus, pushing the instance's confidence to surpass the abstention threshold fails to provide risk improvements. However, such a technique could be employed in scenarios where risk degradation can be tolerated.

4 Analysis

What is the distribution of model's original confidence on the instances that get answered after applying post-abstention method? In Figure 5, we show the distribution of model's original confidence on SQuAD instances that get answered by RETOP at abstention threshold 0.5. Green-colored bars represent the number of instances answered from each confidence bucket. *We found that RETOP answers a large number of instances from the high*

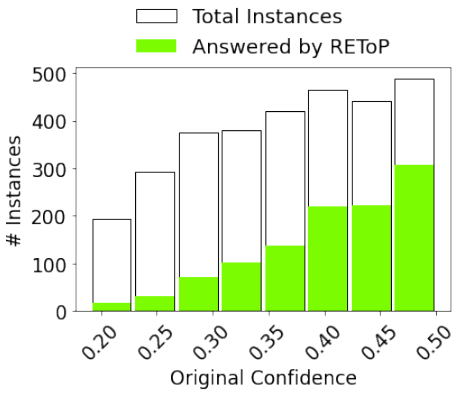

Figure 5: Distribution of QA model's confidence on SQuAD instances that get answered after applying RETOP at abstention threshold 0.5.

confidence buckets; however, instances from even low confidence buckets get answered. This can further be controlled using the weight parameter (α) in the overall confidence computation.

How often do the system's predictions change after applying RETOP and what is its impact? RETOP can either boost the confidence of the top most prediction of the given model or can select a different answer by re-examining its top N predictions. In Figure 6, we specifically analyze the latter scenario i.e. the instances on which RETOP's prediction differs from the original model's prediction. At a threshold of 0.5, the original system abstains on 3411 SQuAD instances and after applying RETOP, it answers 1110 of those instances. Out of these 1110 instances, the RETOP changes the prediction on 186 instances. The original prediction is incorrect in more cases (99 vs 87) and after applying RETOP, the system gives 116 correct predictions and only 70 incorrect. This implies that by overriding the original system's prediction, RETOP improves the system's accuracy. However, in some cases, it also changed a correct prediction to incorrect but such cases are lesser than the former.

To what extent do the instances answered by different post-abstention methods overlap? In Figure 7, we demonstrate the Venn diagram of SQuAD instances answered by RETOP and Ensembling (Mean) approaches at abstention threshold 0.5. RETOP answers 1110 instances while Ensembling answers 277 and there 127 common instances between the two approaches. This indicates that the two sets are not mutually exclusive i.e. there are some instances that get targeted by both the approaches; however, there are a significant number of instances that are not in the intersection. This result motivates studying composite or sequential application of different post-abstention methods to further improve the post-abstention performance.

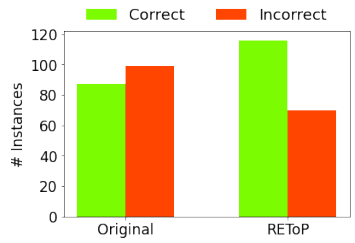

Figure 6: Number of correct (green) and incorrect (red) predictions on those abstained SQuAD instances where RETOP surpasses the abstention threshold of 0.5 but its prediction differs from the original system.

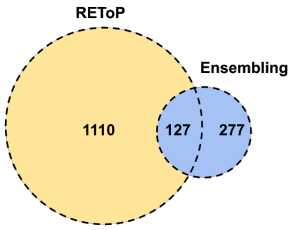

Figure 7: Venn diagram of abstained SQuAD instances answered by RETOP and Ensembling (Mean) approaches at abstention threshold 0.5.

5 Conclusion and Discussion

In this work, we formulated 'Post-Abstention', a task that allows re-attempting the abstained instances of the given selective prediction system with the aim of increasing its *coverage* without significantly sacrificing the *accuracy*. We also explored several baseline methods for this task. Through comprehensive experiments on 11 QA datasets, we showed that these methods lead to considerable performance improvements in both in-domain and out-of-domain settings. We further performed a thorough analysis that resulted in several interesting findings.

Looking forward, we believe that our work opens up several avenues for new research, such as exploring *test-time adaptation*, *knowledge hunting*, and other human intervention techniques like *asking clarification questions* as post-abstention methods (discussed in Appendix D). Studying the impact of composite or sequential application of multiple post-abstention methods in another promising direction. Furthermore, prior selective prediction

methods can also be repurposed and explored for this task. We plan to pursue these crucial research directions in our future work. Finally, we hope our work will encourage further research in this important area and facilitate the development of more reliable NLP systems.

Limitations

The proposed post-abstention methods require additional computation and storage. Despite this additional requirement, we note that this is not a serious concern as current devices have high storage capacity and computation hardware. Furthermore, additional computation for training auxiliary model in REToP is required only once and just an inference is required at evaluation time which has a much lower computation cost. Moreover, the risk mitigation that comes with the post-abstention methods weighs much more than the computational or storage overhead in terms of importance. Secondly, human-intervention techniques require a human to be a participant and contribute in the answering process. However, these approaches do not expect the participating human to be an expert in the task. Like other empirical research, it is difficult to exactly predict the magnitude of improvement a post-abstention method can bring. Our idea of exploring sequential application of multiple post-abstention methods addresses this concern and can be used based on the application requirements.

Acknowledgement

We thank the anonymous reviewers for their insightful feedback. This research was supported by DARPA SAIL-ON program.

References

Mohammad Aliannejadi, Julia Kiseleva, Aleksandr Chuklin, Jeff Dalton, and Mikhail Burtsev. 2020. Convai3: Generating clarifying questions for open-domain dialogue systems (clariq). *arXiv preprint arXiv:2009.11352*.

Mohammad Aliannejadi, Julia Kiseleva, Aleksandr Chuklin, Jeff Dalton, and Mikhail Burtsev. 2021. Building and evaluating open-domain dialogue corpora with clarifying questions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4473–4484, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Pratyay Banerjee, Tejas Gokhale, and Chitta Baral. 2021. Self-supervised test-time learning for reading comprehension. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1200–1211, Online. Association for Computational Linguistics.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*.

Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. 2022. Contrastive test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 295–305.

William B Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. 2017. Searchqa: A new q&a dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*.

Ran El-Yaniv et al. 2010. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(5).

Hady Elsahar and Matthias Gallé. 2019. To annotate or not? predicting performance drop under domain shift. 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)*, pages 2163–2173, Hong Kong, China. Association for Computational Linguistics.

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. MRQA 2019 shared task: Evaluating generalization in reading comprehension. In *Proceedings of 2nd Machine Reading*

for Reading Comprehension (MRQA) Workshop at EMNLP.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

Siddhant Garg and Alessandro Moschitti. 2021. Will this question be answered? question filtering via answer model distillation for efficient question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7329–7346, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dan Hendrycks and Kevin Gimpel. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *Proceedings of International Conference on Learning Representations*.

Shankar Iyer, Nikhil Dandekar, and Kornél Csernai. 2017. First quora dataset release: Question pairs. *data. quora. com.*

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1875–1885, New Orleans, Louisiana. Association for Computational Linguistics.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5684–5696, Online. Association for Computational Linguistics.

Daesik Kim, Seonhoon Kim, and Nojun Kwak. 2019. Textbook question answering with multi-modal context graph understanding and self-supervised open-set comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3568–3584, Florence, Italy. Association for Computational Linguistics.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. 2021. Wilds: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5637–5664. PMLR.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

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

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

Lei Li, Yankai Lin, Deli Chen, Shuhuai Ren, Peng Li, Jie Zhou, and Xu Sun. 2021. CascadeBERT: Accelerating inference of pre-trained language models via calibrated complete models cascade. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 475–486, Punta Cana, Dominican Republic. Association for Computational Linguistics.

John Miller, Karl Krauth, Benjamin Recht, and Ludwig Schmidt. 2020. The effect of natural distribution shift on question answering models. In *International Conference on Machine Learning*, pages 6905–6916. PMLR.

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*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Sudha Rao and Hal Daumé III. 2018. Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2737–2746, Melbourne, Australia. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging NLP models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 856–865, Melbourne, Australia. Association for Computational Linguistics.

Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. 2021. Evaluation examples are not equally informative: How should that change NLP leaderboards? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4486–4503, Online. Association for Computational Linguistics.

Amrita Saha, Rahul Aralikatte, Mitesh M. Khapra, and Karthik Sankaranarayanan. 2018. DuoRC: Towards complex language understanding with paraphrased reading comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1683–1693, Melbourne, Australia. Association for Computational Linguistics.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826.

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*, pages 191–200, Vancouver, Canada. Association for Computational Linguistics.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962*.

Svitlana Vakulenko, Nikos Voskarides, Zhucheng Tu, and Shayne Longpre. 2021. A comparison of question rewriting methods for conversational passage retrieval. In *European Conference on Information Retrieval*, pages 418–424. Springer.

Neeraj Varshney and Chitta Baral. 2022. Model cascading: Towards jointly improving efficiency and accuracy of NLP systems. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11007–11021, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Neeraj Varshney, Man Luo, and Chitta Baral. 2022a. Can open-domain qa reader utilize external knowledge efficiently like humans? *arXiv preprint arXiv:2211.12707*.

Neeraj Varshney, Swaroop Mishra, and Chitta Baral. 2022b. ILDAE: Instance-level difficulty analysis of evaluation data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3412–3425, Dublin, Ireland. Association for Computational Linguistics.

Neeraj Varshney, Swaroop Mishra, and Chitta Baral. 2022c. Investigating selective prediction approaches across several tasks in IID, OOD, and adversarial settings. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1995–2002, Dublin, Ireland. Association for Computational Linguistics.

Neeraj Varshney, Swaroop Mishra, and Chitta Baral. 2022d. Towards improving selective prediction ability of NLP systems. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 221–226, Dublin, Ireland. Association for Computational Linguistics.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing NLP. 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)*, pages 2153–2162, Hong Kong, China. Association for Computational Linguistics.

Xinyi Wang, Yulia Tsvetkov, Sebastian Ruder, and Graham Neubig. 2021. Efficient test time adapter ensembling for low-resource language varieties. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 730–737, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Ji Xin, Raphael Tang, Yaoliang Yu, and Jimmy Lin. 2021. The art of abstention: Selective prediction and error regularization for natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1040–1051, Online. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Hamed Zamani, Susan T. Dumais, Nick Craswell, Paul N. Bennett, and Gord Lueck. 2020a. Generating clarifying questions for information retrieval. *Proceedings of The Web Conference 2020*.

Hamed Zamani, Gord Lueck, Everest Chen, Rodolfo Quispe, Flint Luu, and Nick Craswell. 2020b. Mimics: A large-scale data collection for search clarification. In *Proceedings of the 29th ACM International on Conference on Information and Knowledge Management*, CIKM '20.

Shujian Zhang, Chengyue Gong, and Eunsol Choi. 2021. Knowing more about questions can help: Improving calibration in question answering. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1958–1970, Online. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

Appendix

A Ensembling (Mean) Performance

Table 5 shows the performance of using Ensembling (Mean) as a post-abstention method for a few selected abstention threshold values. For each dataset, we provide three rows: the first row ('*Given*') shows the coverage and risk values of the given selective prediction system at specified abstention thresholds, the second row ('*Ens*') shows the coverage and risk after applying the post-abstention method on the abstained instances of the given selective prediction system, and the final row ('$G@Ens_{cov}$') shows the risk of the given selective system at the coverage achieved by *Ens* method. For the post-abstention method to be effective the risk in the second row should be less than that in the third row and the magnitude of difference corresponds to the improvement. The last column '*Total Risk Improvement*' shows the overall improvement aggregated over all confidence thresholds ranging between 0 and 1 at an interval of 0.02.

B Dataset Statistics

Table 4 shows the statistics of all evaluation datasets used in this work. SQuAD corresponds to the in-domain dataset while the remaining 10 datasets are out-of-domain. We use the pre-processed data from the MRQA shared task (Fisch et al., 2019).

C Differentiating REToP from Calibration

REToP is different from calibration based techniques presented in (Kamath et al., 2020; Varshney et al., 2022c) in the following aspects:
(a) Firstly, REToP does not require a held-out dataset unlike calibration based methods that infer the model on the held-out dataset to gather instances on which the model in incorrect.
(b) Secondly, the auxiliary model trained in REToP predicts the likelihood of correctness of (context, question, prediction) triplet i.e. it is used for each of the top N prediction individually. This is in contrast to calibrators that predicts a single score for an instance and ignores the top N predictions.
(c) Finally, we use the entire context, question, and the prediction to predict its correctness likelihood score unlike feature-based calibrator models in which a random-forest model is trained using just syntax-level features such as length of question,

Dataset	Size	Dataset	Size
SQuAD	10507	HotpotQA	5901
RE	2948	RACE	674
NewsQA	4212	SearchQA	16980
TriviaQA	7785	NQ	12836
DROP	1503	DuoRC	1501
TBQA	1503		

Table 4: Statistics of evaluation data used in this work.

semantic similarity of prediction with the question, etc.

D Other Post-Abstention Techniques

Asking clarifying questions to the user in order to get information about the question has started to received considerable research attention in conversational, web search, and information retrieval settings (Aliannejadi et al., 2021, 2020; Zamani et al., 2020a; Zhang et al., 2020; Zamani et al., 2020b). These techniques can be leveraged/adapted for the post-abstention task.

Test-time adaptation is another promising research area in which the model is adapted at test-time depending on the instance. This is being studied in both computer vision (Chen et al., 2022) and language processing (Wang et al., 2021; Banerjee et al., 2021).

Cascading systems in which stronger and stronger models are conditionally used for inference is also an interesting avenue to explore with respect to Post-Abstention (Varshney and Baral, 2022; Li et al., 2021; Varshney et al., 2022a).

E Coverage 100% for Human Intervention Methods

We believe that the ability to identify situations when there is no good answer in the top N returned candidates is a very difficult task (for the humans also) and it requires even more cognitive skills than just selecting the best answer from the provided answer candidates. Because of this reason, the coverage is 100%.

F Comparison with Other Selective Prediction Methods

In this work, we presented a new QA setting and studied the performance of several baseline methods for this task. The focus of this work is on studying the risk improvement that can be achieved in this problem setup. We consciously do not pitch the approaches for this task as competitors of the existing selective prediction approaches. In fact, these approaches are **complimentary** to the selective prediction approaches. A post-abstention method can be used with any selective prediction method as the first step.

Dataset	Model	0.2		0.32		0.36		0.48		0.54		0.60		0.68		Total Risk
		Cov↑	Risk↓	Cov↑	Risk↓	Cov↑	Risk↓	Cov↑	Risk↓	Cov↑	Risk↓	Cov↑	Risk↓	Cov↑	Risk↓	**Improvement↑**
	Given (G)	96.65	32.45	87.24	28.10	83.34	26.69	69.94	21.91	62.57	19.91	56.23	17.98	47.92	15.43	
SQuAD	Ens	97.64	32.88	89.51	28.93	87.64	28.24	72.46	22.71	65.12	20.58	58.37	18.7	49.59	15.89	
(in-domain)	G@Ens$_{cov}$	-	32.96	-	29.09	-	28.26	-	22.58	-	20.65	-	18.66	-	15.91	0.29
	Given (G)	97.54	67.65	89.56	65.88	85.39	65.13	71.75	62.71	64.77	61.56	58.19	60.34	49.25	58.29	
HotpotQA	Ens	98.59	67.84	91.93	66.23	90.41	65.92	75.65	63.17	68.45	62.22	61.31	60.72	52.26	58.88	
	G@Ens$_{cov}$	-	67.9	-	66.37	-	66.04	-	63.4	-	62.14	-	60.91	-	58.94	0.93
	Given (G)	97.59	44.49	89.01	40.51	85.41	39.04	74.08	34.16	66.86	30.54	60.58	27.94	54.10	24.20	
RE	Ens	98.27	44.56	92.2	41.35	90.57	40.71	77.44	34.87	70.86	31.45	64.86	29.08	56.07	24.74	
	G@Ens$_{cov}$	-	44.82	-	42.27	-	41.42	-	35.58	-	32.47	-	30.02	-	25.54	21.72
	Given (G)	89.02	80.5	71.07	77.04	66.17	75.56	51.34	72.54	43.47	69.62	36.2	68.85	29.97	63.86	
RACE	Ens	91.69	80.42	73.89	77.71	71.51	77.18	53.71	72.65	46.88	70.25	40.21	69.0	31.6	64.79	
	G@Ens$_{cov}$	-	80.88	-	77.31	-	77.13	-	72.93	-	71.43	-	70.11	-	65.09	16.72
	Given (G)	93.90	69.76	80.91	66.40	75.5	64.91	60.30	60.79	53.30	58.8	47.17	56.62	39.32	54.11	
NewsQA	Ens	95.56	70.24	83.52	67.14	81.13	66.49	63.01	61.53	55.75	59.45	49.53	57.19	41.17	54.21	
	G@Ens$_{cov}$	-	70.18	-	67.02	-	66.46	-	61.63	-	59.67	-	57.33	-	54.67	11.92
	Given (G)	96.15	86.68	81.77	85.67	75.77	85.34	58.64	84.08	50.22	83.58	42.67	83.33	34.46	82.55	
SearchQA	Ens	98.0	86.82	87.31	85.79	84.7	85.61	65.65	84.1	56.86	83.65	48.46	83.16	38.73	82.36	
	G@Ens$_{cov}$	-	86.83	-	86.05	-	85.87	-	84.52	-	84.03	-	83.59	-	82.94	17.05
	Given (G)	96.67	67.31	86.89	65.05	82.54	63.82	68.81	60.39	61.44	58.39	55.11	56.48	47.12	54.03	
TriviaQA	Ens	98.01	67.58	89.88	65.71	87.99	65.15	72.31	60.95	65.0	59.13	58.47	56.9	49.67	54.38	
	G@Ens$_{cov}$	-	67.64	-	65.76	-	65.3	-	61.38	-	59.25	-	57.55	-	54.94	9.5
	Given (G)	92.37	63.78	79.04	59.99	74.87	58.77	60.60	53.51	54.03	51.00	47.94	48.31	41.70	45.27	
NQ	Ens	94.59	64.35	83.46	60.82	81.32	60.16	64.83	54.7	58.05	52.17	51.8	49.8	44.33	46.31	
	G@Ens$_{cov}$	-	64.43	-	61.31	-	60.79	-	55.03	-	52.61	-	50.01	-	46.82	13.4
	Given (G)	95.74	88.46	81.17	87.38	76.11	87.33	62.34	86.23	53.69	85.38	48.77	84.45	43.05	85.01	
DROP	Ens	97.6	88.48	85.63	87.72	83.17	87.28	65.34	86.15	56.55	85.65	50.37	84.54	44.78	84.99	
	G@Ens$_{cov}$	-	88.47	-	87.72	-	87.52	-	86.05	-	85.63	-	84.54	-	84.84	1.57
	Given (G)	97.20	68.68	87.87	66.41	84.21	65.82	71.09	62.42	64.16	61.47	57.16	59.91	50.03	58.46	
DuoRC	Ens	98.0	68.86	90.34	67.11	88.61	66.84	73.82	63.36	66.96	62.19	59.96	60.78	51.57	58.4	
	Original@cov	-	68.91	-	67.18	-	66.69	-	63.18	-	61.79	-	60.07	-	58.91	-1.69
	Given (G)	94.34	67.14	80.9	63.32	75.65	61.92	57.49	56.02	49.63	52.14	41.45	51.04	34.07	50.00	
TBQA	Ens	95.94	67.55	84.3	64.17	81.1	63.33	62.28	56.94	53.96	54.25	45.78	52.33	37.72	51.15	
	G@Ens$_{cov}$	-	67.45	-	64.33	-	63.38	-	57.05	-	54.38	-	52.03	-	50.53	-6.93

Table 5: **Performance of Ensembling (Mean) as a post-abstention method** for selected abstention thresholds. The QA model is trained using SQuAD training data and evaluated on SQuAD (in-domain) and 10 out-of-domain datasets. For each dataset, we provide three rows: first row ('*Given*') shows the coverage and risk values of the given selective prediction system at different abstention thresholds, second row ('*Ens*') shows the coverage and risk after applying Ens on abstained instances of the given system, and third row ('$G@Ens_{cov}$') shows risk of the given system at the coverage achieved by *Ens*. For the post abstention method to be effective, risk in the second row should be less than that in the third row and the magnitude of difference corresponds to the improvement. The last column corresponds to the overall improvement aggregated over all confidences ranging from 0 to 1 at an interval of 0.02. ↓ and ↑ indicate that lower (risk) and higher (coverage, risk improvement) values are better respectively.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
We have Limitations Section at the end of the paper after Conclusion

☐ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Introduction

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

References

☑ B1. Did you cite the creators of artifacts you used?
We use the publicly available standard NLP datasets in this work with appropriate citations and references.

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
We do not create any artifcats in this reserach. We use the publicly available standard NLP datasets in this work with proper citations and references.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
We do not create any artifcats in this reserach. We use the publicly available standard NLP datasets in this work with proper citations and references.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
We do not collect any data for this research and use standard publicly available NLP datasets

☒ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
We do not collect any data for this research

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 4

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

C ☑ Did you run computational experiments?

Sections 3 and 4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Sections 3 and 4

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Sections 3

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Sections 3 and 4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Sections 3 and 4

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.