Logic-Guided Data Augmentation and Regularization for Consistent Question Answering

Akari Asai[†] and **Hannaneh Hajishirzi**[†‡]
†University of Washington ‡Allen Institute for AI
{akari, hannaneh}@cs.washington.edu

MRC dataset

- To create comparison questions that require inferential knowledge and reasoning ability, annotators need to understand context presented in multiple paragraphs or carefully ground a question to the given situation.

- Our method leverages **logical and linguistic knowledge** to augment labeled training data and then uses a **consistency-based regularizer** to train the model.

Inconsistent predictions

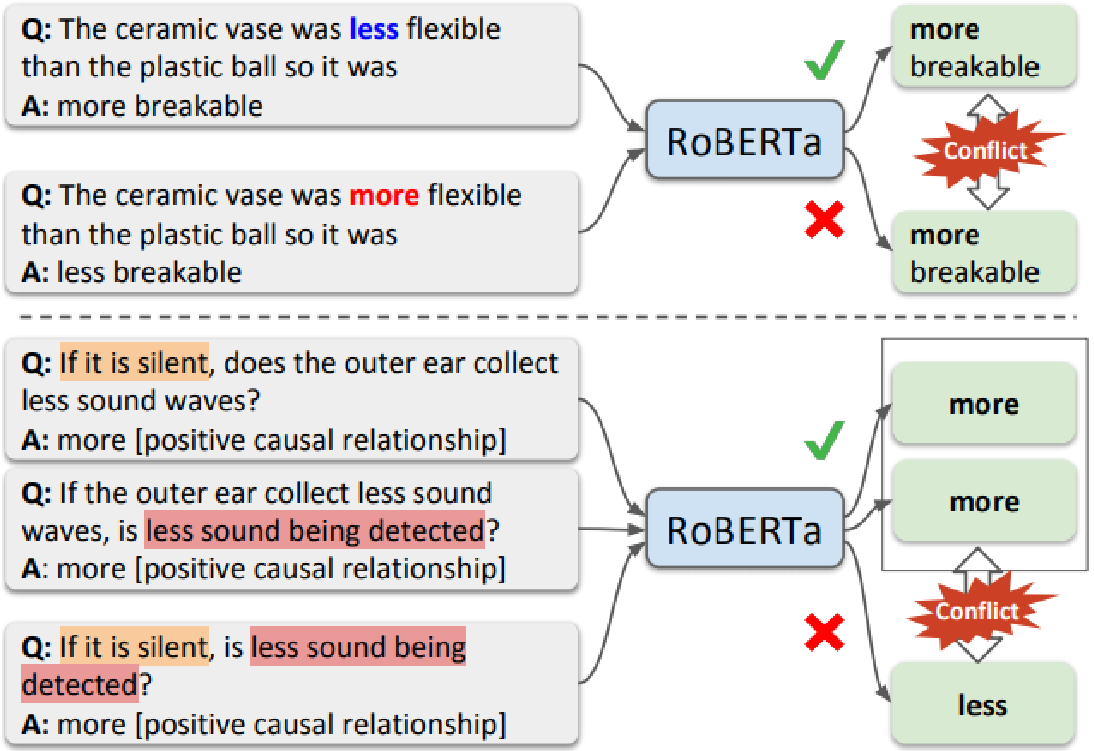
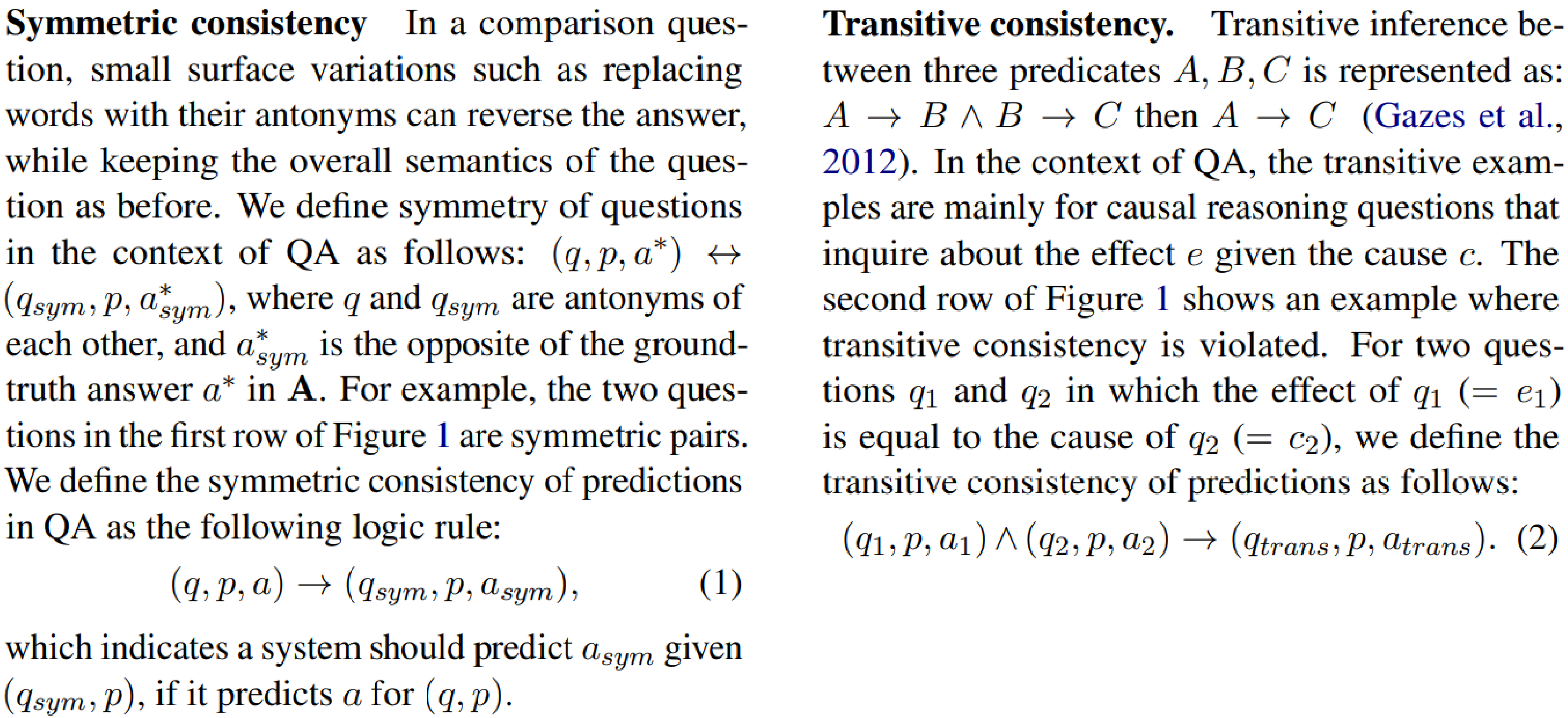

Q: The ceramic vase was less flexible than the plastic ball so it was
A: more breakable
Q: The ceramic vase was more flexible than the plastic ball so it was
A: less breakable
RoBERTa
✓
more breakable
Conflict
✗
more breakable
Q: If it is silent, does the outer ear collect less sound waves?
A: more [positive causal relationship]
Q: If the outer ear collect less sound waves, is less sound being detected?
A: more [positive causal relationship]
Q: If it is silent, is less sound being detected?
A: more [positive causal relationship]
RoBERTa
✓
more
more
Conflict
✗
less

Figure 1: Inconsistent predictions by RoBERTa. Top row shows an example of symmetric inconsistency and the second row shows an example of transitive inconsistency. The examples are partially modified.

Consistent Question Answering

- Let (q, p, a) be a question, a paragraph and an answer predicted by a model. A is a set of answer candidates. Each element of A can be a span in p, a class category, or an arbitrary answer choice. X = {q, p, a} represents a logic atom.

Consistent Question Answering



Symmetric consistency In a comparison question, small surface variations such as replacing words with their antonyms can reverse the answer, while keeping the overall semantics of the question as before. We define symmetry of questions in the context of QA as follows: $(q, p, a^*) \leftrightarrow (q_{sym}, p, a^*_{sym})$, where q and q_{sym} are antonyms of each other, and a^*_{sym} is the opposite of the ground-truth answer a^* in $\mathbf{A}$. For example, the two questions in the first row of Figure 1 are symmetric pairs. We define the symmetric consistency of predictions in QA as the following logic rule:

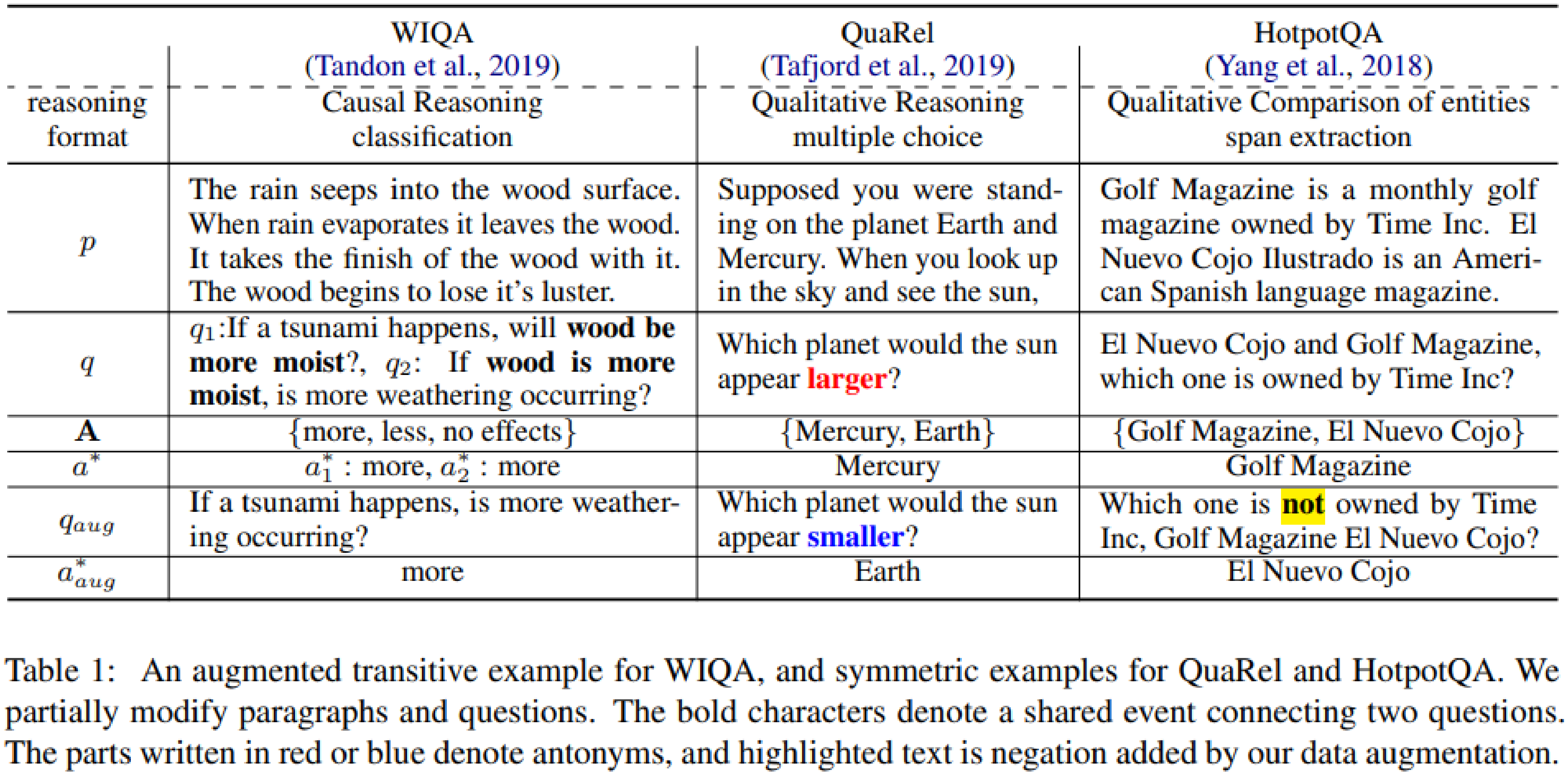
$$(q, p, a) \rightarrow (q_{sym}, p, a_{sym}), \tag{1}$$

which indicates a system should predict a_{sym} given (q_{sym}, p), if it predicts a for (q, p).

Transitive consistency. Transitive inference between three predicates A, B, C is represented as: $A \rightarrow B \wedge B \rightarrow C$ then $A \rightarrow C$ (Gazes et al., 2012). In the context of QA, the transitive examples are mainly for causal reasoning questions that inquire about the effect e given the cause c. The second row of Figure 1 shows an example where transitive consistency is violated. For two questions q_1 and q_2 in which the effect of q_1 $(= e_1)$ is equal to the cause of q_2 $(= c_2)$, we define the transitive consistency of predictions as follows:

$$(q_1, p, a_1) \wedge (q_2, p, a_2) \rightarrow (q_{trans}, p, a_{trans}). \tag{2}$$

Method



	WIQA (Tandon et al., 2019)	QuaRel (Tafjord et al., 2019)	HotpotQA (Yang et al., 2018)
reasoning format	Causal Reasoning classification	Qualitative Reasoning multiple choice	Qualitative Comparison of entities span extraction
p	The rain seeps into the wood surface. When rain evaporates it leaves the wood. It takes the finish of the wood with it. The wood begins to lose it's luster.	Supposed you were standing on the planet Earth and Mercury. When you look up in the sky and see the sun,	Golf Magazine is a monthly golf magazine owned by Time Inc. El Nuevo Cojo Ilustrado is an American Spanish language magazine.
q	q_1:If a tsunami happens, will **wood be more moist**?, q_2: If **wood is more moist**, is more weathering occurring?	Which planet would the sun appear **larger**?	El Nuevo Cojo and Golf Magazine, which one is owned by Time Inc?
A	{more, less, no effects}	{Mercury, Earth}	{Golf Magazine, El Nuevo Cojo}
a^*	a_1^* : more, a_2^* : more	Mercury	Golf Magazine
q_{aug}	If a tsunami happens, is more weathering occurring?	Which planet would the sun appear **smaller**?	Which one is **not** owned by Time Inc, Golf Magazine El Nuevo Cojo?
a_{aug}^*	more	Earth	El Nuevo Cojo

Table 1: An augmented transitive example for WIQA, and symmetric examples for QuaRel and HotpotQA. We partially modify paragraphs and questions. The bold characters denote a shared event connecting two questions. The parts written in red or blue denote antonyms, and highlighted text is negation added by our data augmentation.

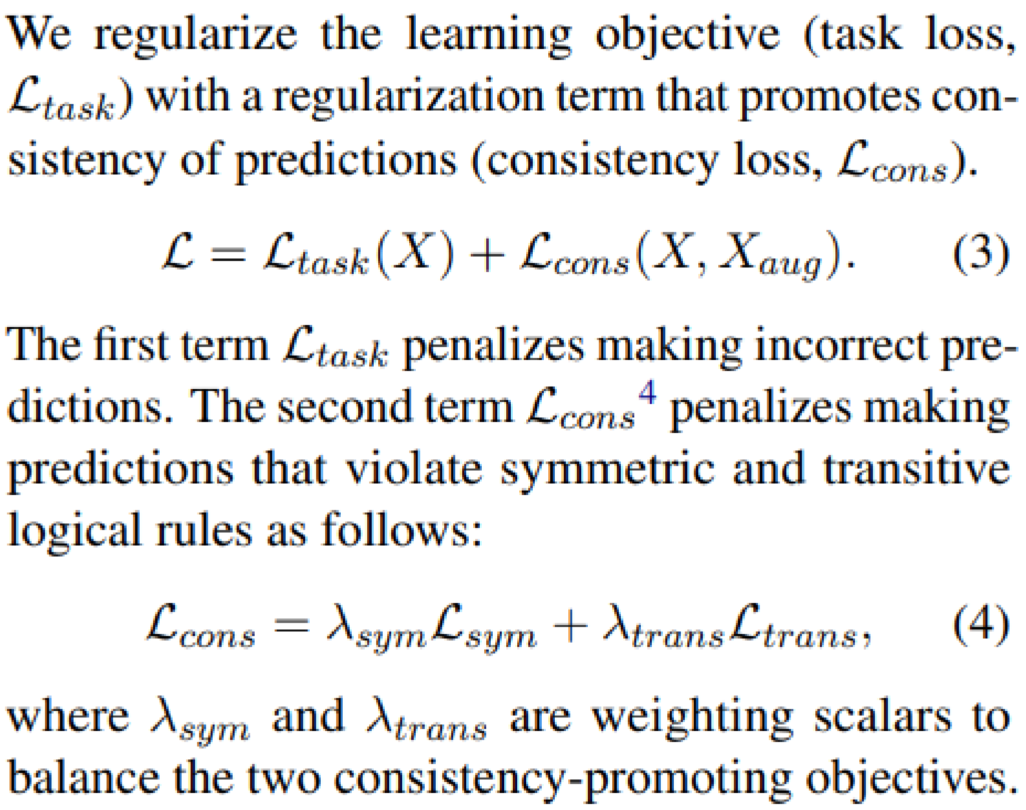
Augmenting symmetric examples

- replace words with their antonyms
    - we select top frequent **adjectives or verbs** with polarity (e.g., smaller, increases) from training corpora, and expert annotators write antonyms for each of the frequent words (we denote this small dictionary as D)
- add or remove words.
    - we add negation words or remove negation words (e.g., not). For all of the questions in training data, if a question includes a word in D for the operation (a), or matches a template (e.g., which * is <-> which * is not)

Augmenting transitive examples

- We first find a pair of two cause-effect questions X_1 = (q_1, p, $a*_1$) and X_2 = (q_2, p, $a*_2$), whose q_1 and q_2 consist of (c_1, e_1) and (c_2, e_2), where e_1 = c_2 holds. When $a*_1$ is a positive causal relationship, we create a new example X_{trans} = (q_3, p, $a*_2$) for q_3 = (c_1, e_2).

Logic-guided Consistency Regularization



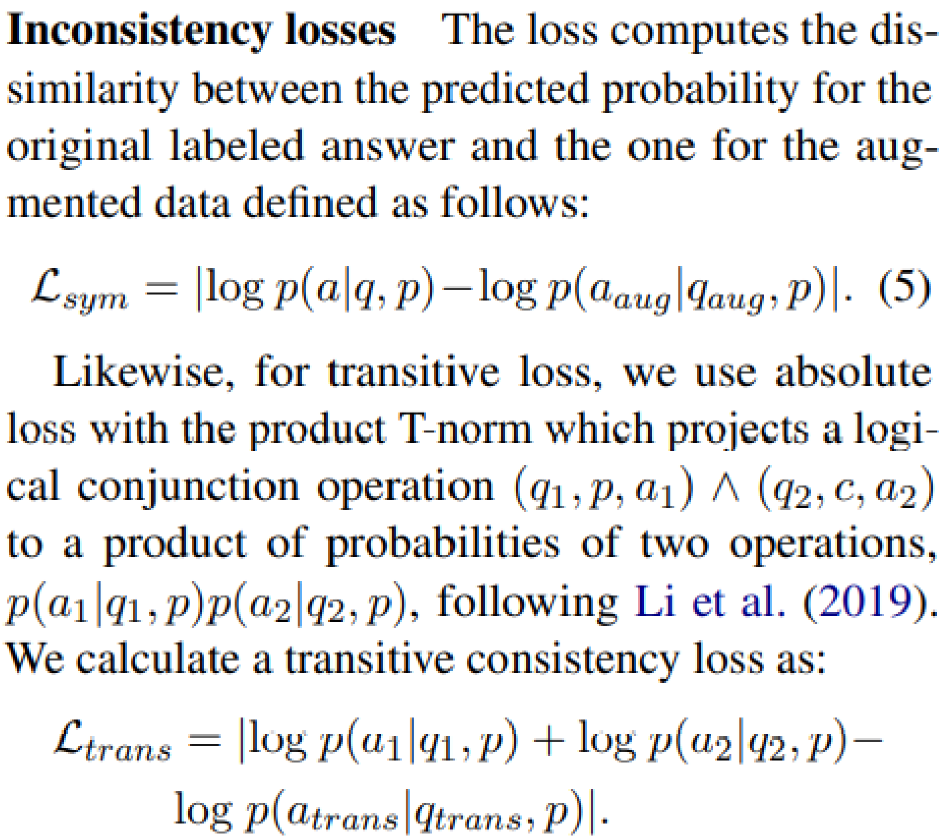
We regularize the learning objective (task loss, $\mathcal{L}_{task}$) with a regularization term that promotes consistency of predictions (consistency loss, $\mathcal{L}_{cons}$).

$$\mathcal{L} = \mathcal{L}_{task}(X) + \mathcal{L}_{cons}(X, X_{aug}). \qquad (3)$$

The first term $\mathcal{L}_{task}$ penalizes making incorrect predictions. The second term $\mathcal{L}_{cons}$[4] penalizes making predictions that violate symmetric and transitive logical rules as follows:

$$\mathcal{L}_{cons} = \lambda_{sym}\mathcal{L}_{sym} + \lambda_{trans}\mathcal{L}_{trans}, \qquad (4)$$

where λ_{sym} and λ_{trans} are weighting scalars to balance the two consistency-promoting objectives.



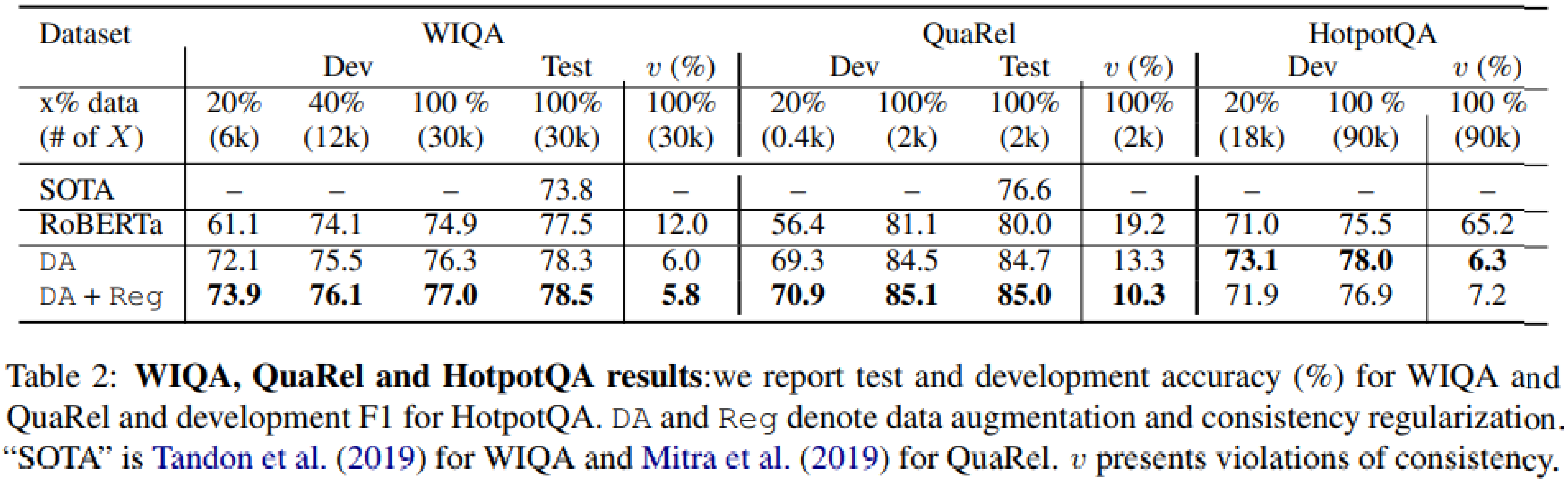
Inconsistency losses The loss computes the dissimilarity between the predicted probability for the original labeled answer and the one for the augmented data defined as follows:

$$\mathcal{L}_{sym} = |\log p(a|q,p) - \log p(a_{aug}|q_{aug},p)|. \quad (5)$$

Likewise, for transitive loss, we use absolute loss with the product T-norm which projects a logical conjunction operation $(q_1, p, a_1) \wedge (q_2, c, a_2)$ to a product of probabilities of two operations, $p(a_1|q_1,p)p(a_2|q_2,p)$, following Li et al. (2019). We calculate a transitive consistency loss as:

$$\mathcal{L}_{trans} = |\log p(a_1|q_1,p) + \log p(a_2|q_2,p) - \\ \log p(a_{trans}|q_{trans},p)|.$$

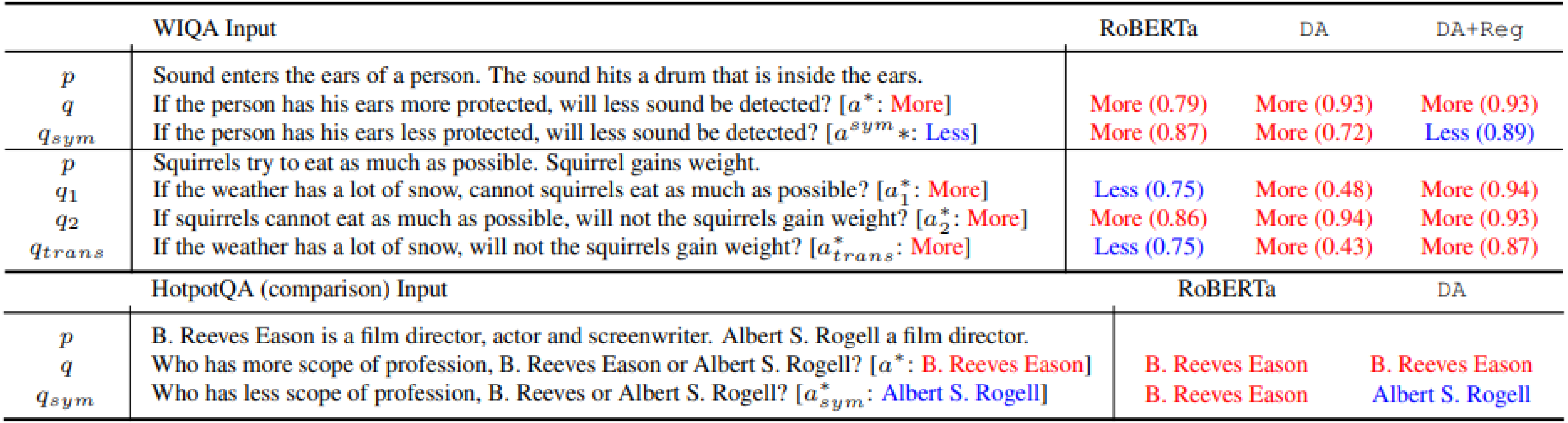
Experiments



Dataset	WIQA					QuaRel				HotpotQA		
	Dev			Test	v (%)	Dev		Test	v (%)	Dev		v (%)
x% data (# of X)	20% (6k)	40% (12k)	100 % (30k)	100% (30k)	100% (30k)	20% (0.4k)	100% (2k)	100% (2k)	100% (2k)	20% (18k)	100 % (90k)	100 % (90k)
SOTA	–	–	–	73.8	–	–	–	76.6	–	–	–	–
RoBERTa	61.1	74.1	74.9	77.5	12.0	56.4	81.1	80.0	19.2	71.0	75.5	65.2
DA	72.1	75.5	76.3	78.3	6.0	69.3	84.5	84.7	13.3	**73.1**	**78.0**	**6.3**
DA + Reg	**73.9**	**76.1**	**77.0**	**78.5**	**5.8**	**70.9**	**85.1**	**85.0**	**10.3**	71.9	76.9	7.2

Table 2: **WIQA, QuaRel and HotpotQA results**:we report test and development accuracy (%) for WIQA and QuaRel and development F1 for HotpotQA. DA and Reg denote data augmentation and consistency regularization. "SOTA" is Tandon et al. (2019) for WIQA and Mitra et al. (2019) for QuaRel. v presents violations of consistency.

Experiments

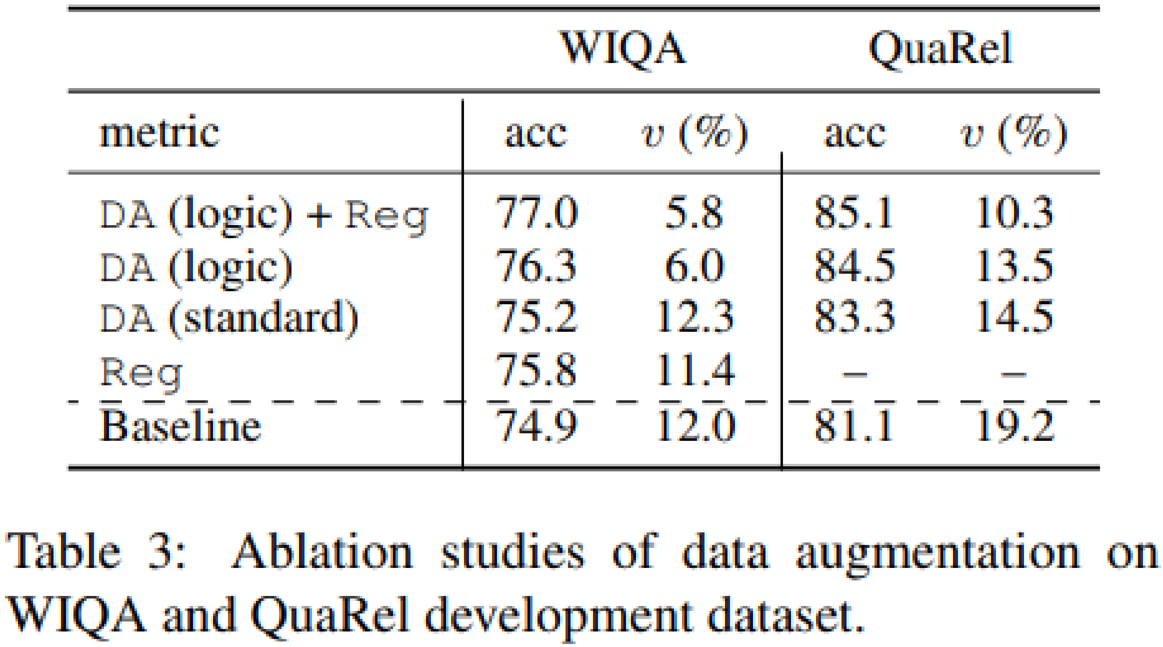


	WIQA Input	RoBERTa	DA	DA+Reg
p	Sound enters the ears of a person. The sound hits a drum that is inside the ears.			
q	If the person has his ears more protected, will less sound be detected? [a^*: More]	More (0.79)	More (0.93)	More (0.93)
q_{sym}	If the person has his ears less protected, will less sound be detected? [$a^{sym}*$: Less]	More (0.87)	More (0.72)	Less (0.89)
p	Squirrels try to eat as much as possible. Squirrel gains weight.			
q_1	If the weather has a lot of snow, cannot squirrels eat as much as possible? [a_1^*: More]	Less (0.75)	More (0.48)	More (0.94)
q_2	If squirrels cannot eat as much as possible, will not the squirrels gain weight? [a_2^*: More]	More (0.86)	More (0.94)	More (0.93)
q_{trans}	If the weather has a lot of snow, will not the squirrels gain weight? [a_{trans}^*: More]	Less (0.75)	More (0.43)	More (0.87)

	HotpotQA (comparison) Input	RoBERTa	DA
p	B. Reeves Eason is a film director, actor and screenwriter. Albert S. Rogell a film director.		
q	Who has more scope of profession, B. Reeves Eason or Albert S. Rogell? [a^*: B. Reeves Eason]	B. Reeves Eason	B. Reeves Eason
q_{sym}	Who has less scope of profession, B. Reeves or Albert S. Rogell? [a_{sym}^*: Albert S. Rogell]	B. Reeves Eason	Albert S. Rogell

Table 4: Qualitative comparison of RoBERTa, + DA, + DA + Reg. The examples are partially modified.

Experiments



	WIQA		QuaRel	
metric	acc	v (%)	acc	v (%)
DA (logic) + Reg	77.0	5.8	85.1	10.3
DA (logic)	76.3	6.0	84.5	13.5
DA (standard)	75.2	12.3	83.3	14.5
Reg	75.8	11.4	–	–
Baseline	74.9	12.0	81.1	19.2

Table 3: Ablation studies of data augmentation on WIQA and QuaRel development dataset.

No Answer is Better Than Wrong Answer: A Reflection Model for Document Level Machine Reading Comprehension

Xuguang Wang[1] **Linjun Shou**[1] **Ming Gong**[1] **Nan Duan**[2] **Daxin Jiang**[1‡]

[1]Microsoft STCA NLP Group, Beijing, China

[2]Microsoft Research Asia, Beijing, China

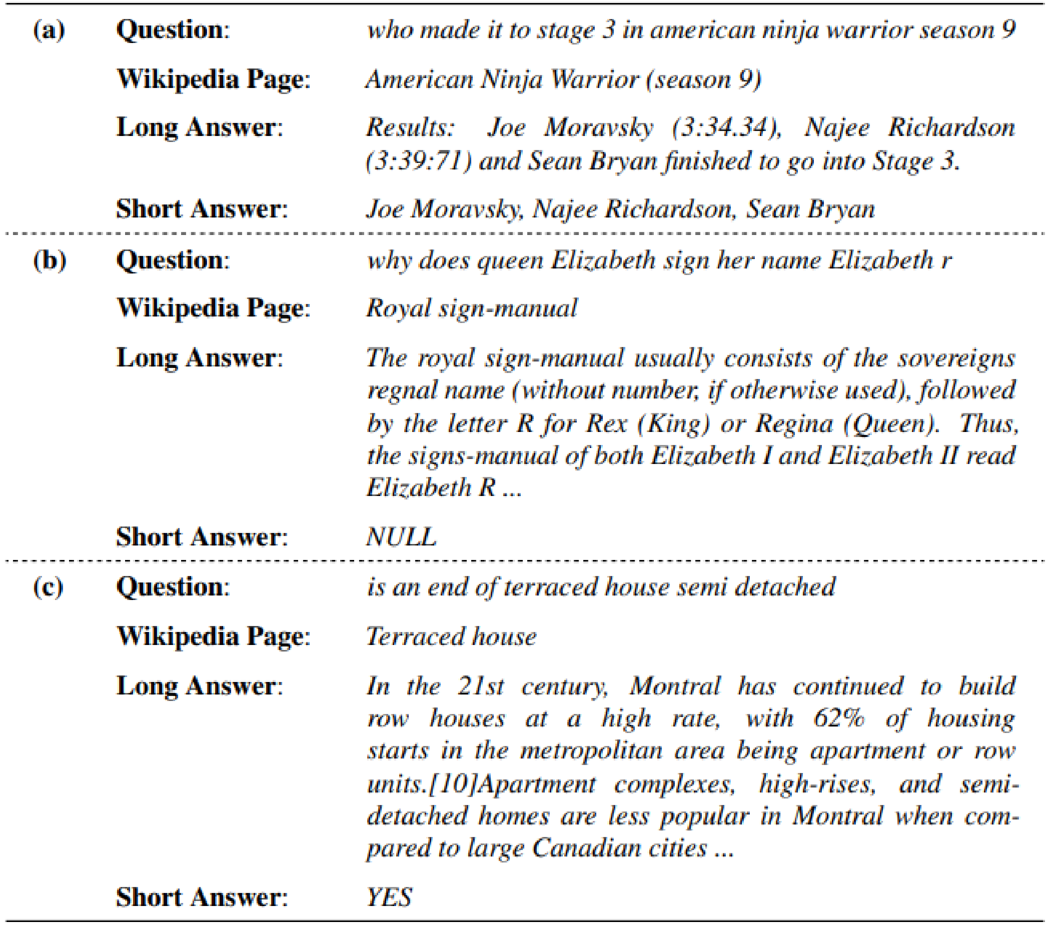
{xugwang,lisho,migon,nanduan,djiang}@microsoft.com

Natural Questions (NQ) benchmark

- NQ’s answers are not only at different levels of granularity (long and short), but also of richer types (including no-answer, yes/no, single-span and multi-span)

- There are richer answer types in the NQ task. In addition to indicating textual answer spans (long and short), the models need to handle cases including no-answer (51%), multi-span short answer (3.5%), and yes/no (1%) answer.

NQ challenge



(a)	**Question**:	*who made it to stage 3 in american ninja warrior season 9*
	Wikipedia Page:	*American Ninja Warrior (season 9)*
	Long Answer:	*Results: Joe Moravsky (3:34.34), Najee Richardson (3:39:71) and Sean Bryan finished to go into Stage 3.*
	Short Answer:	*Joe Moravsky, Najee Richardson, Sean Bryan*
(b)	**Question**:	*why does queen Elizabeth sign her name Elizabeth r*
	Wikipedia Page:	*Royal sign-manual*
	Long Answer:	*The royal sign-manual usually consists of the sovereigns regnal name (without number, if otherwise used), followed by the letter R for Rex (King) or Regina (Queen). Thus, the signs-manual of both Elizabeth I and Elizabeth II read Elizabeth R ...*
	Short Answer:	*NULL*
(c)	**Question**:	*is an end of terraced house semi detached*
	Wikipedia Page:	*Terraced house*
	Long Answer:	*In the 21st century, Montral has continued to build row houses at a high rate, with 62% of housing starts in the metropolitan area being apartment or row units.[10]Apartment complexes, high-rises, and semi-detached homes are less popular in Montral when compared to large Canadian cities ...*
	Short Answer:	*YES*

Table 1: Example of NQ challenge, short answer cases: (a) Multi-span answer, (b) No-answer, (c) Yes/No.

Method

- We first train an all types handling MRC model.
- Then, we leverage the trained MRC model to inference all the training data, train a second model, called the **Reflection model** takes as inputs the predicted answer, its context and MRC head features to predict a more accurate confidence score which distinguish the right answer from the wrong ones.

Model

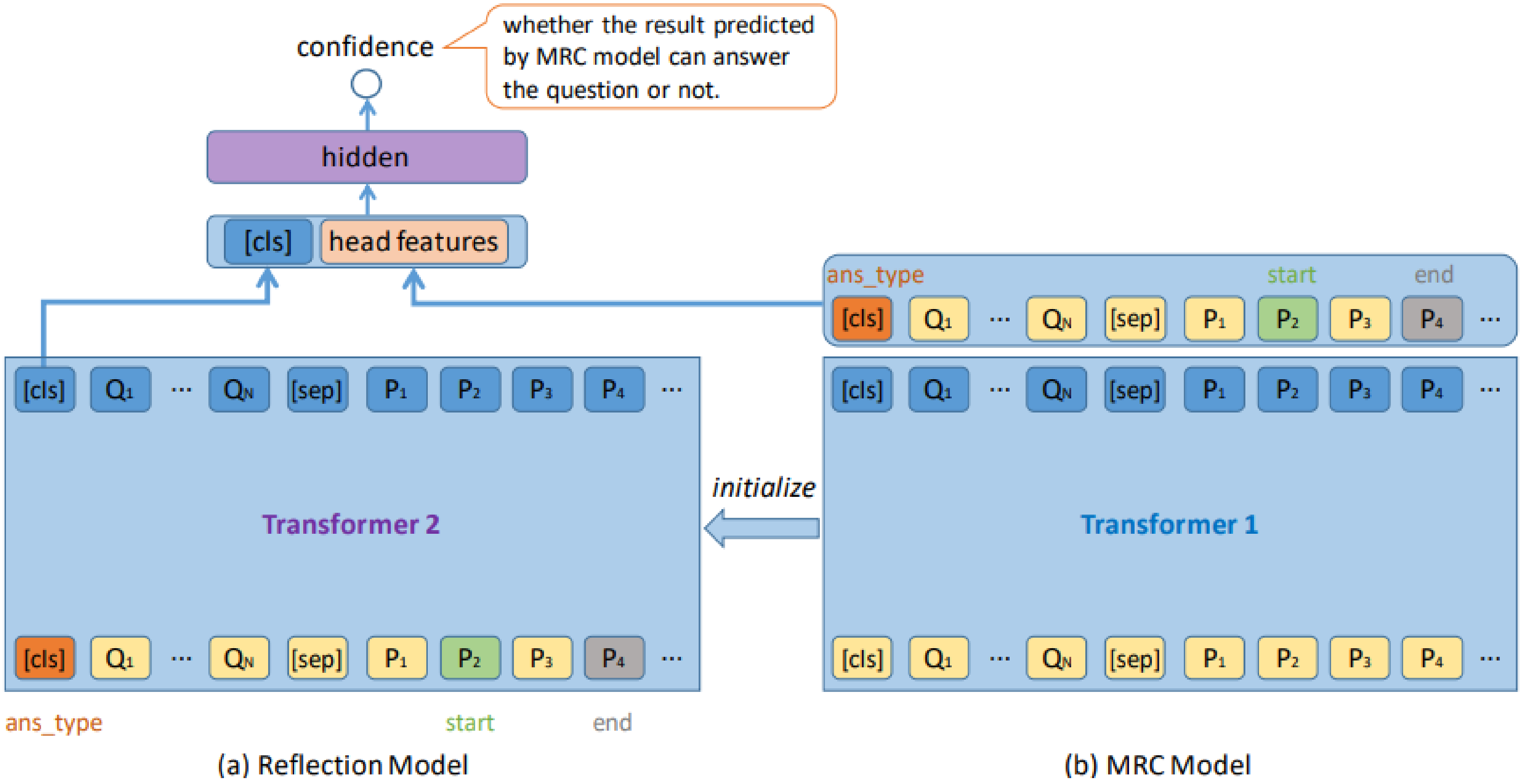
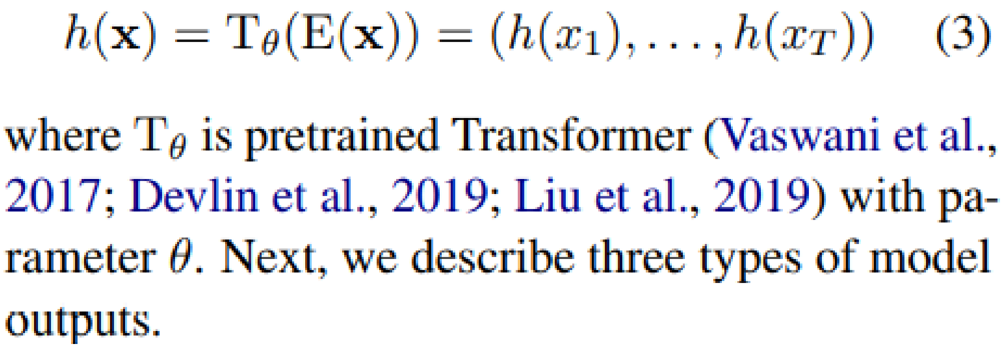


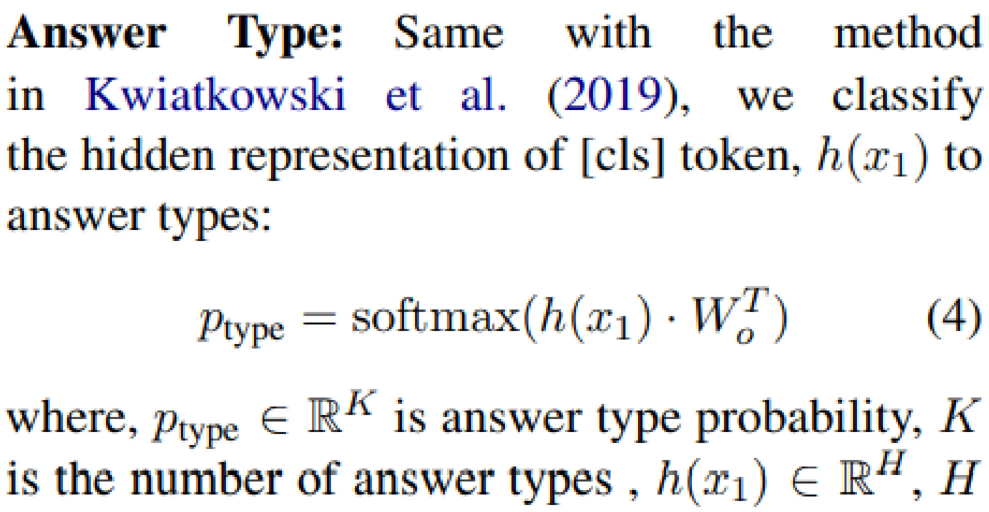
Figure 1: Overview of our proposed *Reflection Net*, consisting of MRC model and its corresponding Reflection model. MRC model try its best to predict answer, Reflection model output corresponding answer confidence score. The left arrow denotes when training, Reflection model is initialized with the parameters of trained MRC model.

MRC Model



$$h(\mathbf{x}) = \mathrm{T}_\theta(\mathrm{E}(\mathbf{x})) = (h(x_1), \ldots, h(x_T)) \quad (3)$$

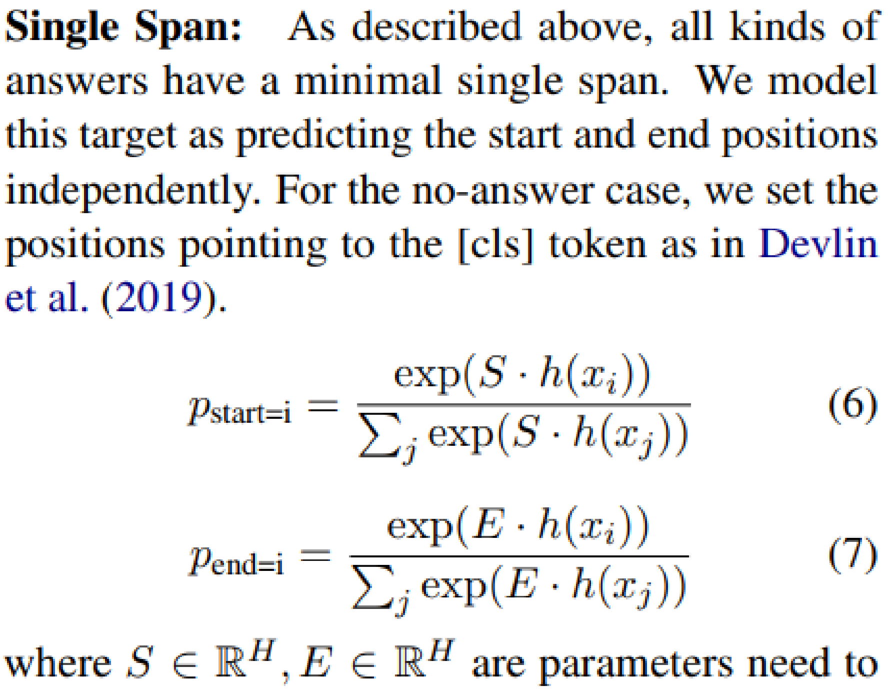
where T_θ is pretrained Transformer (Vaswani et al., 2017; Devlin et al., 2019; Liu et al., 2019) with parameter θ. Next, we describe three types of model outputs.



Answer Type: Same with the method in Kwiatkowski et al. (2019), we classify the hidden representation of [cls] token, $h(x_1)$ to answer types:

$$p_{\text{type}} = \text{softmax}(h(x_1) \cdot W_o^T) \quad (4)$$

where, $p_{\text{type}} \in \mathbb{R}^K$ is answer type probability, K is the number of answer types , $h(x_1) \in \mathbb{R}^H$, H

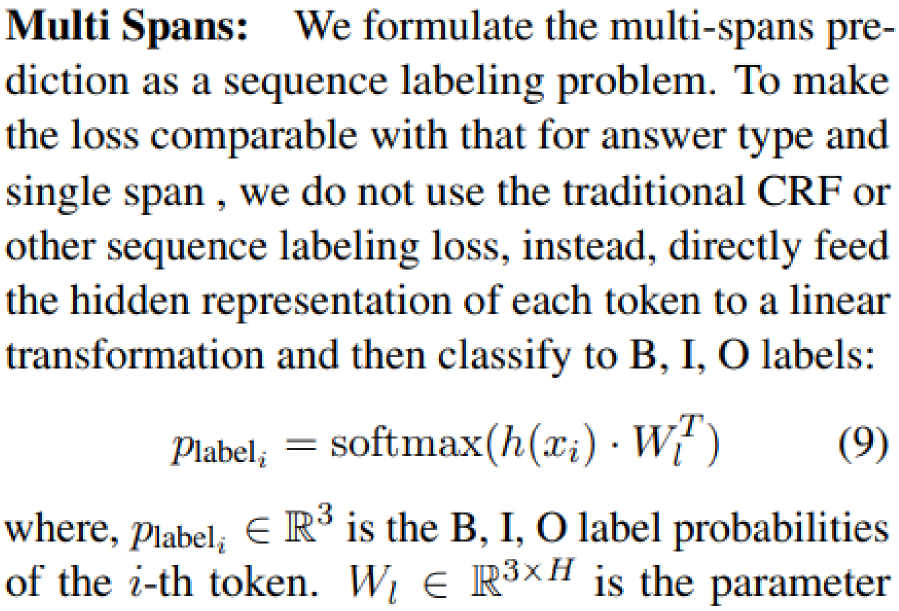


Single Span: As described above, all kinds of answers have a minimal single span. We model this target as predicting the start and end positions independently. For the no-answer case, we set the positions pointing to the [cls] token as in Devlin et al. (2019).

$$p_{\text{start=i}} = \frac{\exp(S \cdot h(x_i))}{\sum_j \exp(S \cdot h(x_j))} \quad (6)$$

$$p_{\text{end=i}} = \frac{\exp(E \cdot h(x_i))}{\sum_j \exp(E \cdot h(x_j))} \quad (7)$$

where $S \in \mathbb{R}^H, E \in \mathbb{R}^H$ are parameters need to



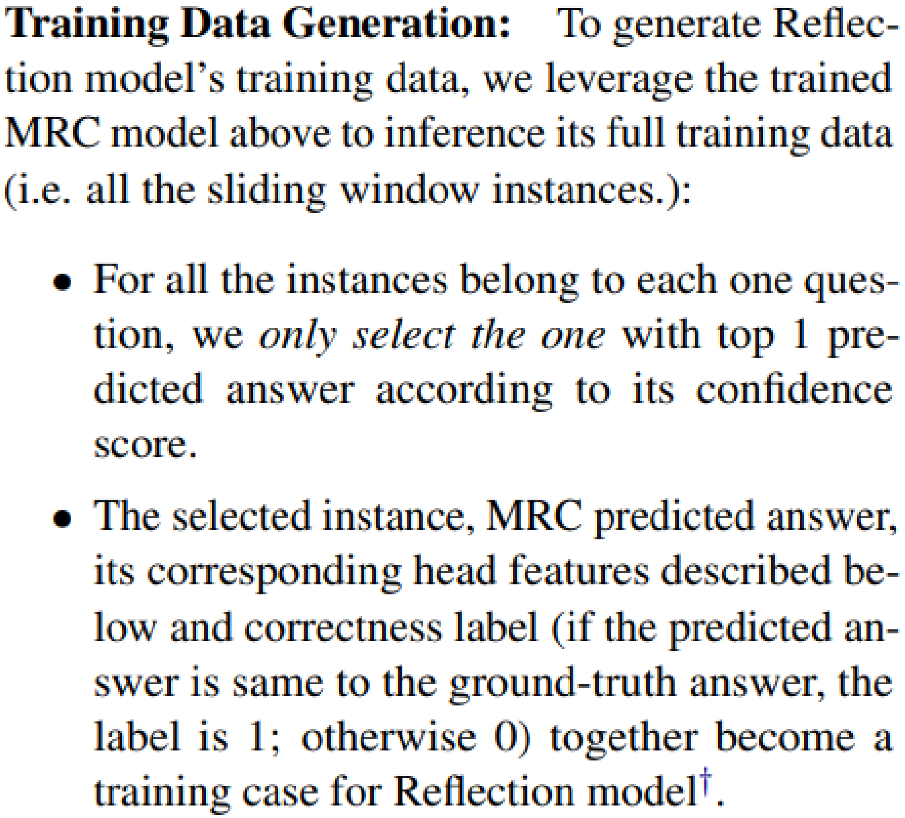
Multi Spans: We formulate the multi-spans prediction as a sequence labeling problem. To make the loss comparable with that for answer type and single span , we do not use the traditional CRF or other sequence labeling loss, instead, directly feed the hidden representation of each token to a linear transformation and then classify to B, I, O labels:

$$p_{\text{label}_i} = \text{softmax}(h(x_i) \cdot W_l^T) \quad (9)$$

where, $p_{\text{label}_i} \in \mathbb{R}^3$ is the B, I, O label probabilities of the i-th token. $W_l \in \mathbb{R}^{3\times H}$ is the parameter

Reflection Model

- Reflection model target a more precise confidence score which distinguish the right answer from two kinds of wrong ones.



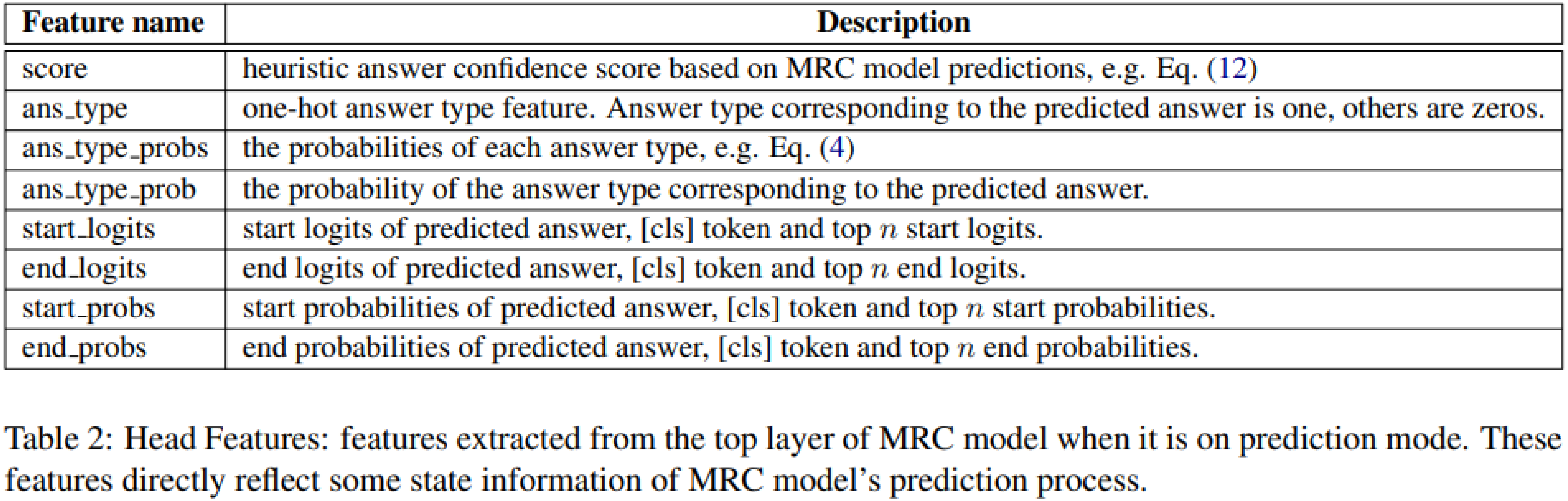
Training Data Generation: To generate Reflection model's training data, we leverage the trained MRC model above to inference its full training data (i.e. all the sliding window instances.):

- For all the instances belong to each one question, we *only select the one* with top 1 predicted answer according to its confidence score.
- The selected instance, MRC predicted answer, its corresponding head features described below and correctness label (if the predicted answer is same to the ground-truth answer, the label is 1; otherwise 0) together become a training case for Reflection model[†].

Model Training

- Initialize Reflection model with the parameters of the trained MRC model
- To directly receive important state information of the MRC model, we extract **head features** from the top layer of the MRC model when it is predicting the answer.
- The head features are concatenated with the hidden representation of [cls] token, then followed by a hidden layer for final confidence prediction.

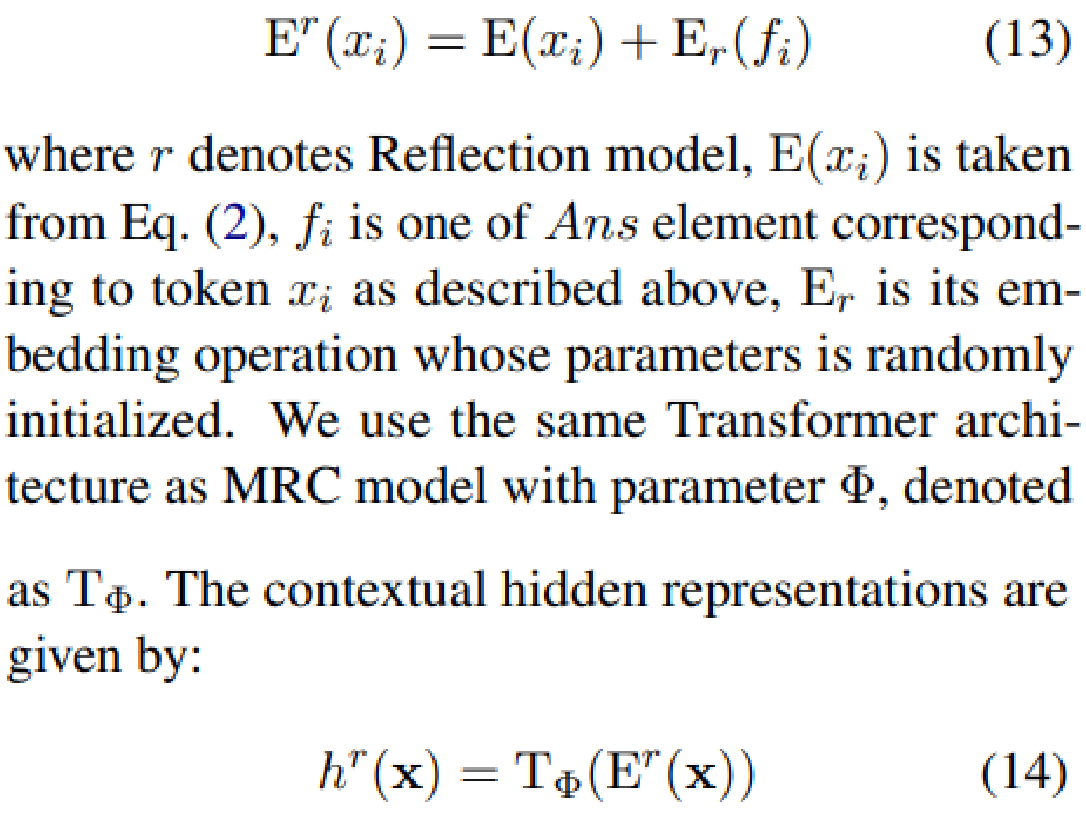
Head Features



Feature name	Description
score	heuristic answer confidence score based on MRC model predictions, e.g. Eq. (12)
ans_type	one-hot answer type feature. Answer type corresponding to the predicted answer is one, others are zeros.
ans_type_probs	the probabilities of each answer type, e.g. Eq. (4)
ans_type_prob	the probability of the answer type corresponding to the predicted answer.
start_logits	start logits of predicted answer, [cls] token and top n start logits.
end_logits	end logits of predicted answer, [cls] token and top n end logits.
start_probs	start probabilities of predicted answer, [cls] token and top n start probabilities.
end_probs	end probabilities of predicted answer, [cls] token and top n end probabilities.

Table 2: Head Features: features extracted from the top layer of MRC model when it is on prediction mode. These features directly reflect some state information of MRC model's prediction process.

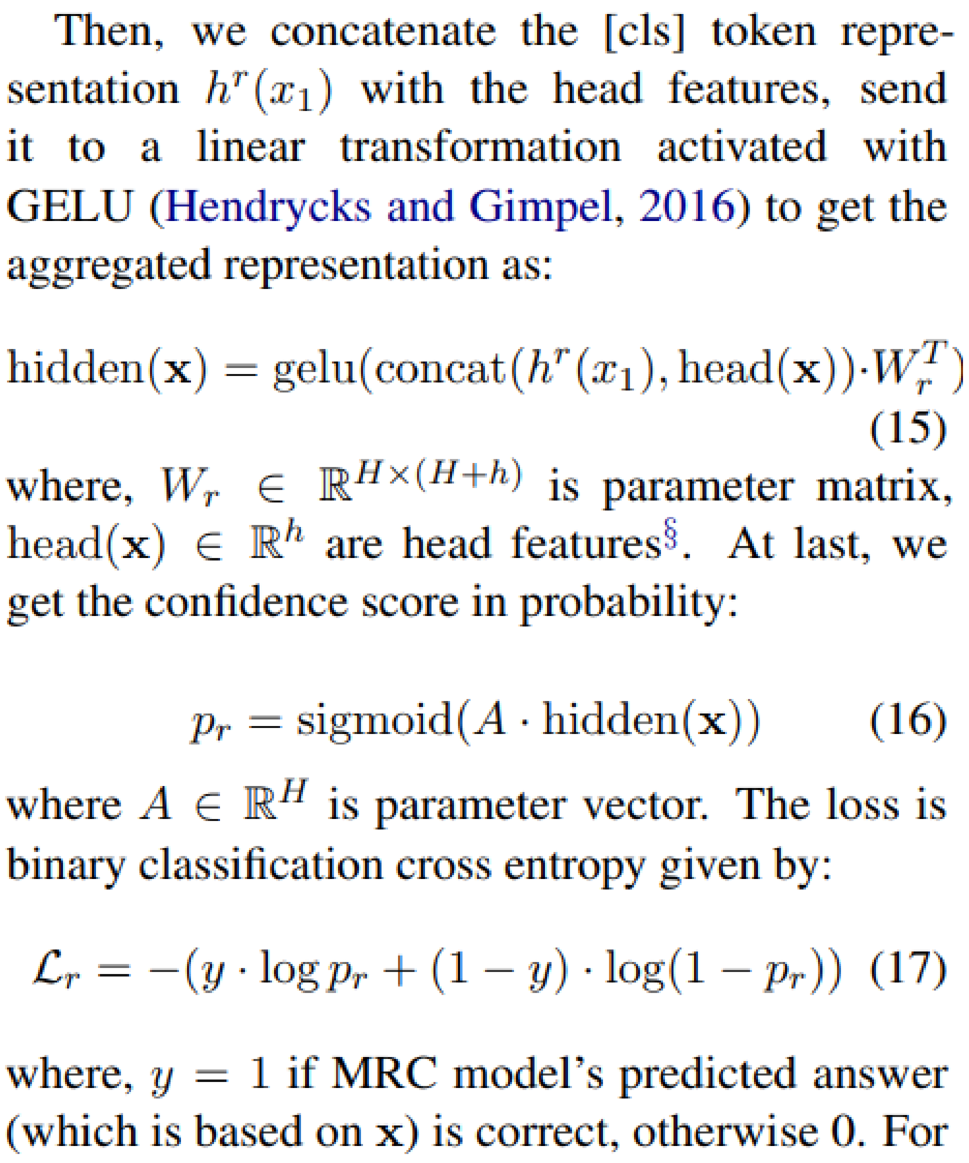
Model Training



$$\mathrm{E}^r(x_i) = \mathrm{E}(x_i) + \mathrm{E}_r(f_i) \tag{13}$$

where r denotes Reflection model, $\mathrm{E}(x_i)$ is taken from Eq. (2), f_i is one of Ans element corresponding to token x_i as described above, E_r is its embedding operation whose parameters is randomly initialized. We use the same Transformer architecture as MRC model with parameter Φ, denoted as T_Φ. The contextual hidden representations are given by:

$$h^r(\mathbf{x}) = \mathrm{T}_\Phi(\mathrm{E}^r(\mathbf{x})) \tag{14}$$



Then, we concatenate the [cls] token representation $h^r(x_1)$ with the head features, send it to a linear transformation activated with GELU (Hendrycks and Gimpel, 2016) to get the aggregated representation as:

$$\mathrm{hidden}(\mathbf{x}) = \mathrm{gelu}(\mathrm{concat}(h^r(x_1), \mathrm{head}(\mathbf{x}))\cdot W_r^T) \tag{15}$$

where, $W_r \in \mathbb{R}^{H\times(H+h)}$ is parameter matrix, $\mathrm{head}(\mathbf{x}) \in \mathbb{R}^h$ are head features[§]. At last, we get the confidence score in probability:

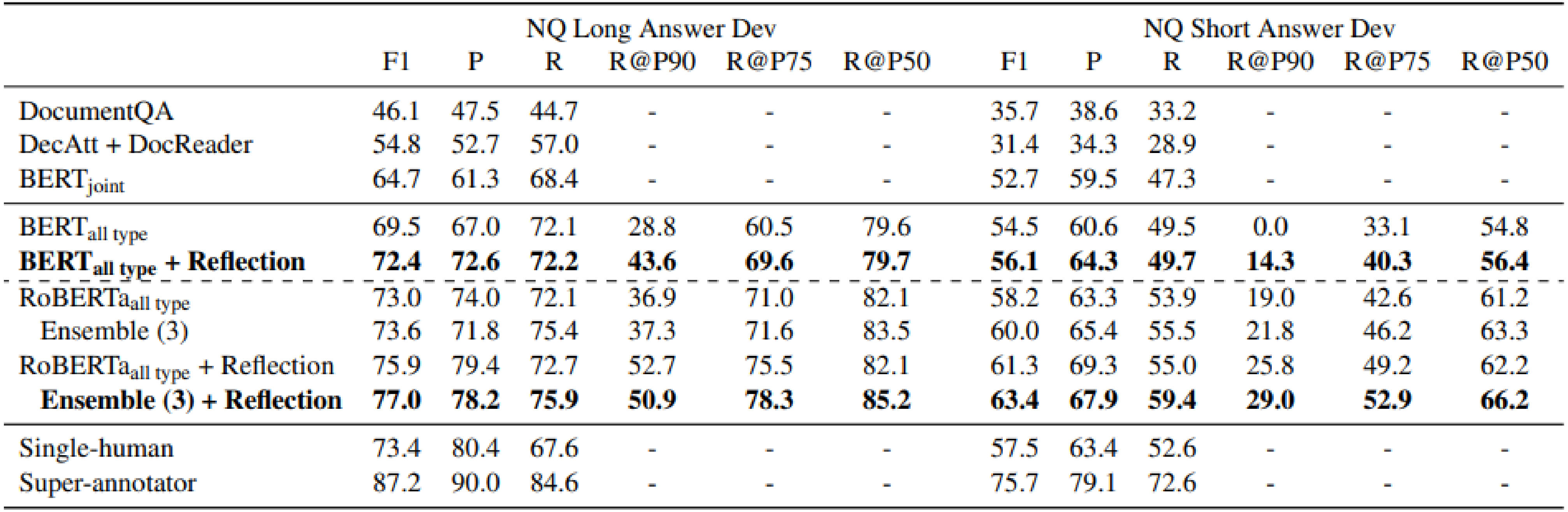
$$p_r = \mathrm{sigmoid}(A \cdot \mathrm{hidden}(\mathbf{x})) \tag{16}$$

where $A \in \mathbb{R}^H$ is parameter vector. The loss is binary classification cross entropy given by:

$$\mathcal{L}_r = -(y \cdot \log p_r + (1-y)\cdot \log(1-p_r)) \tag{17}$$

where, $y = 1$ if MRC model's predicted answer (which is based on $\mathbf{x}$) is correct, otherwise 0. For

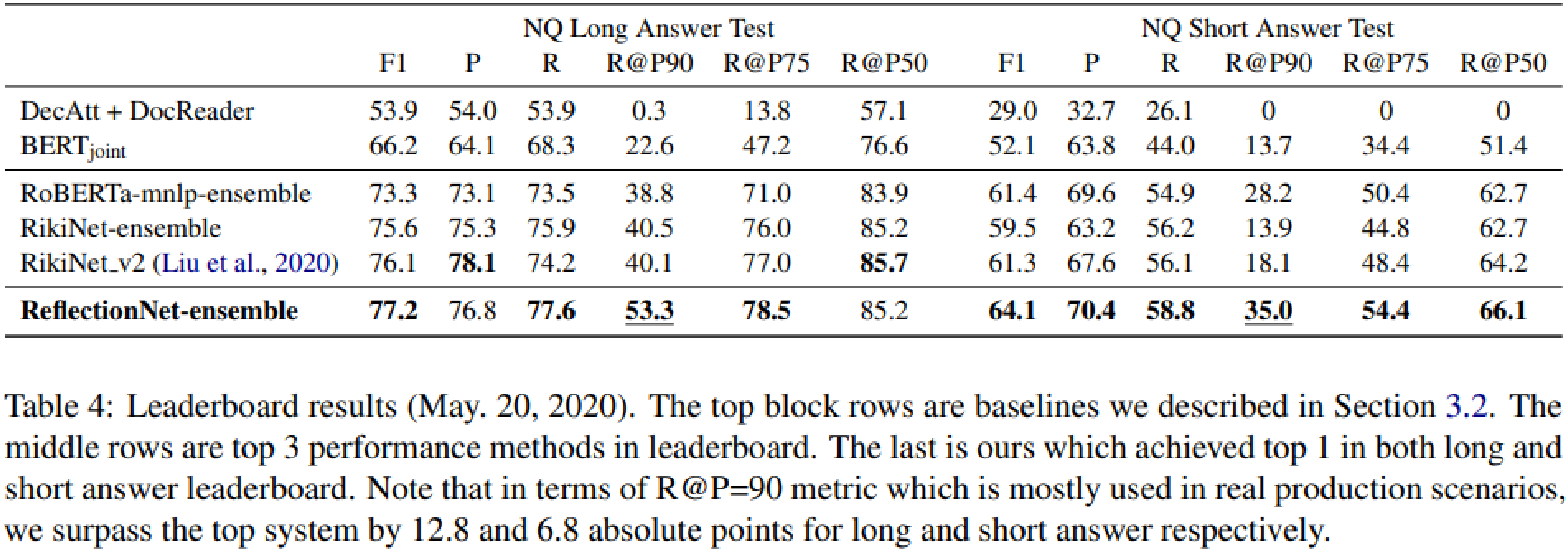
Experiment



	NQ Long Answer Dev						NQ Short Answer Dev					
	F1	P	R	R@P90	R@P75	R@P50	F1	P	R	R@P90	R@P75	R@P50
DocumentQA	46.1	47.5	44.7	-	-	-	35.7	38.6	33.2	-	-	-
DecAtt + DocReader	54.8	52.7	57.0	-	-	-	31.4	34.3	28.9	-	-	-
BERT$_{\text{joint}}$	64.7	61.3	68.4	-	-	-	52.7	59.5	47.3	-	-	-
BERT$_{\text{all type}}$	69.5	67.0	72.1	28.8	60.5	79.6	54.5	60.6	49.5	0.0	33.1	54.8
BERT$_{\text{all type}}$ + Reflection	**72.4**	**72.6**	**72.2**	**43.6**	**69.6**	**79.7**	**56.1**	**64.3**	**49.7**	**14.3**	**40.3**	**56.4**
RoBERTa$_{\text{all type}}$	73.0	74.0	72.1	36.9	71.0	82.1	58.2	63.3	53.9	19.0	42.6	61.2
Ensemble (3)	73.6	71.8	75.4	37.3	71.6	83.5	60.0	65.4	55.5	21.8	46.2	63.3
RoBERTa$_{\text{all type}}$ + Reflection	75.9	79.4	72.7	52.7	75.5	82.1	61.3	69.3	55.0	25.8	49.2	62.2
Ensemble (3) + Reflection	**77.0**	**78.2**	**75.9**	**50.9**	**78.3**	**85.2**	**63.4**	**67.9**	**59.4**	**29.0**	**52.9**	**66.2**
Single-human	73.4	80.4	67.6	-	-	-	57.5	63.4	52.6	-	-	-
Super-annotator	87.2	90.0	84.6	-	-	-	75.7	79.1	72.6	-	-	-

Table 3: NQ development set results. The top block rows are baselines we borrow from Alberti et al. (2019). The last block rows are single human annotator and an ensemble of human annotators. The middle block are ours where BERT$_{\text{all type}}$ and RoBERTa$_{\text{all type}}$ are our MRC model. "+ Reflection" means that our Reflection model is used to provide answer confidence score. Ensemble (3) are three RoBERTa$_{\text{all type}}$ models.

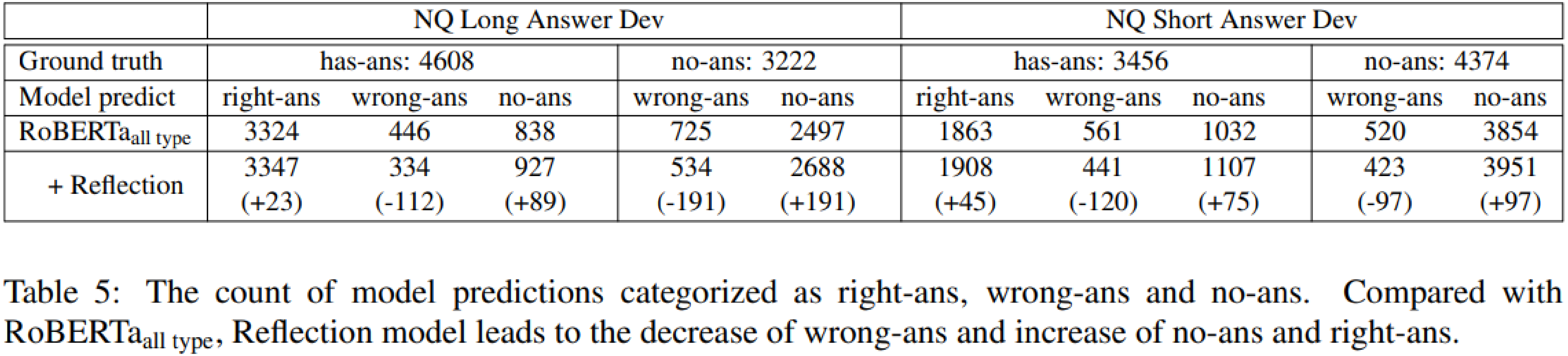
Experiment



	NQ Long Answer Test						NQ Short Answer Test					
	F1	P	R	R@P90	R@P75	R@P50	F1	P	R	R@P90	R@P75	R@P50
DecAtt + DocReader	53.9	54.0	53.9	0.3	13.8	57.1	29.0	32.7	26.1	0	0	0
$\text{BERT}_{\text{joint}}$	66.2	64.1	68.3	22.6	47.2	76.6	52.1	63.8	44.0	13.7	34.4	51.4
RoBERTa-mnlp-ensemble	73.3	73.1	73.5	38.8	71.0	83.9	61.4	69.6	54.9	28.2	50.4	62.7
RikiNet-ensemble	75.6	75.3	75.9	40.5	76.0	85.2	59.5	63.2	56.2	13.9	44.8	62.7
RikiNet_v2 (Liu et al., 2020)	76.1	**78.1**	74.2	40.1	77.0	**85.7**	61.3	67.6	56.1	18.1	48.4	64.2
ReflectionNet-ensemble	**77.2**	76.8	**77.6**	**53.3**	**78.5**	85.2	**64.1**	**70.4**	**58.8**	**35.0**	**54.4**	**66.1**

Table 4: Leaderboard results (May. 20, 2020). The top block rows are baselines we described in Section 3.2. The middle rows are top 3 performance methods in leaderboard. The last is ours which achieved top 1 in both long and short answer leaderboard. Note that in terms of R@P=90 metric which is mostly used in real production scenarios, we surpass the top system by 12.8 and 6.8 absolute points for long and short answer respectively.

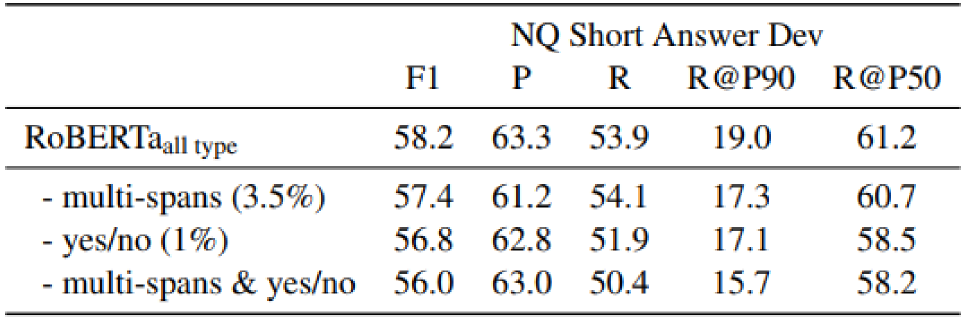
Experiment



	NQ Long Answer Dev					NQ Short Answer Dev				
Ground truth	has-ans: 4608			no-ans: 3222		has-ans: 3456			no-ans: 4374	
Model predict	right-ans	wrong-ans	no-ans	wrong-ans	no-ans	right-ans	wrong-ans	no-ans	wrong-ans	no-ans
RoBERTa$_{\text{all type}}$	3324	446	838	725	2497	1863	561	1032	520	3854
+ Reflection	3347 (+23)	334 (-112)	927 (+89)	534 (-191)	2688 (+191)	1908 (+45)	441 (-120)	1107 (+75)	423 (-97)	3951 (+97)

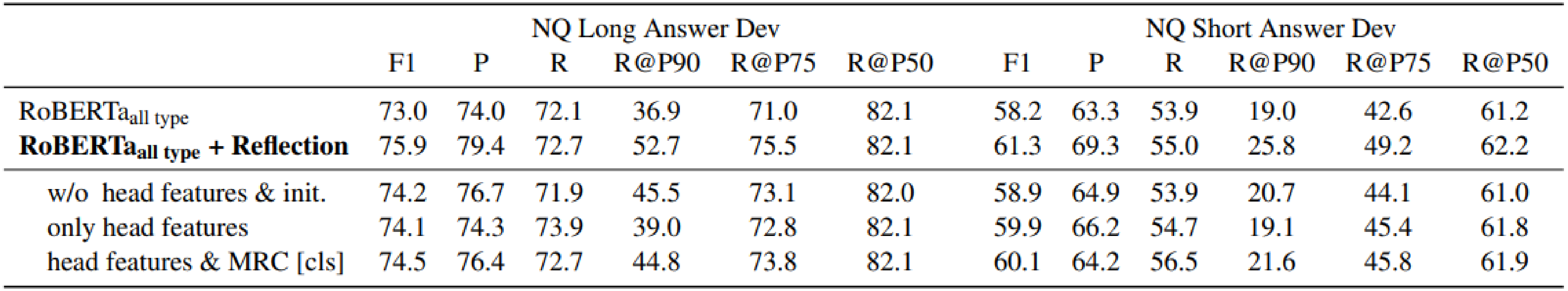
Table 5: The count of model predictions categorized as right-ans, wrong-ans and no-ans. Compared with RoBERTa$_{\text{all type}}$, Reflection model leads to the decrease of wrong-ans and increase of no-ans and right-ans.

Experiment



	NQ Short Answer Dev				
	F1	P	R	R@P90	R@P50
RoBERTa$_{\text{all type}}$	58.2	63.3	53.9	19.0	61.2
- multi-spans (3.5%)	57.4	61.2	54.1	17.3	60.7
- yes/no (1%)	56.8	62.8	51.9	17.1	58.5
- multi-spans & yes/no	56.0	63.0	50.4	15.7	58.2

Table 6: Ablation study on answer types. We compare all answer types handling model with ablation of multi-spans, yes/no type and both.



	NQ Long Answer Dev						NQ Short Answer Dev					
	F1	P	R	R@P90	R@P75	R@P50	F1	P	R	R@P90	R@P75	R@P50
RoBERTa$_{\text{all type}}$	73.0	74.0	72.1	36.9	71.0	82.1	58.2	63.3	53.9	19.0	42.6	61.2
RoBERTa$_{\text{all type}}$ + Reflection	75.9	79.4	72.7	52.7	75.5	82.1	61.3	69.3	55.0	25.8	49.2	62.2
w/o head features & init.	74.2	76.7	71.9	45.5	73.1	82.0	58.9	64.9	53.9	20.7	44.1	61.0
only head features	74.1	74.3	73.9	39.0	72.8	82.1	59.9	66.2	54.7	19.1	45.4	61.8
head features & MRC [cls]	74.5	76.4	72.7	44.8	73.8	82.1	60.1	64.2	56.5	21.6	45.8	61.9

Table 7: Ablation and Variant of Reflection model. There are absence of head features and initialized from MRC model, simple three layer feedforward neural networks which take as input only head features, and lastly, head features integrated with MRC [cls] hidden representation.