Less is More: Rejecting Unreliable Reviews for Product Question Answering

Shiwei Zhang[1,3], Xiuzhen Zhang[1]✉, Jey Han Lau[2], Jeffrey Chan[1], and Cecile Paris[3]

[1] RMIT University, Australia {firstname.lastname}@rmit.edu.au
[2] The University of Melbourne, Australia jeyhan.lau@gmail.com
[3] CSIRO Data61, Australia cecile.paris@data61.csiro.au

Outline

Introduction of Our Reliable PQA Framework

PQA Models

Rejection Model (Confidence Function and Risk Function)

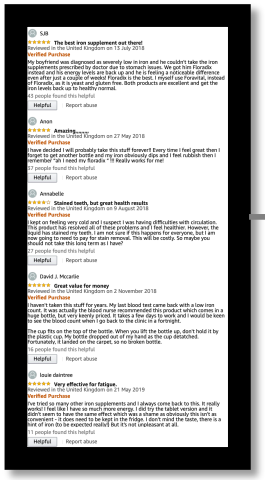
Experiment and Results

Conclusion

If you take this as directed twice a day how long does this size bottle last?

Question

Answer



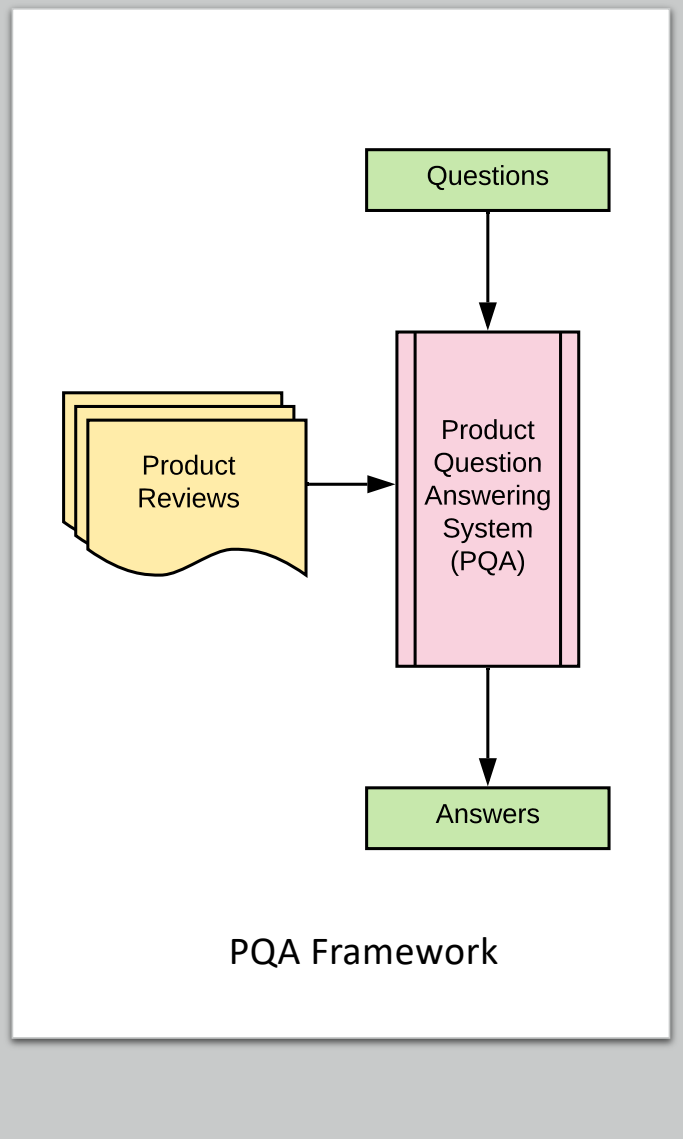
Ranking

Product Question Answering System (PQA)

Top N

1. Following the directions of taking it twice a day, it may last a whole month.
……

PQA Example



PQA Framework

Table 1. Example of an answerable (Q1) and an unanswerable question (Q2). Green denotes high probability/confidence scores, and red otherwise.

Q1: What is the chain for on the side?

Top 3 Ranked Reviews	Prob
- This was driving me crazy but i see that another reviewer explained that grill has wire clip on chain to be used as extended match holder for igniting the gas if the spark mechanism fails to work or is worn out as sometimes happens with any gas grill.	0.99
- PS Could not figure out the purpose of that little chain with the clip attached to the outside of the grill even after reading entire manual.	0.95
- It is to replace an old portable that I have been using for about 10 years.'	0.91

Q2: Does this Dell Inspiron 14R i14RMT-7475s come with dell's warranty?

Top 3 Ranked Reviews	Prob
- I don't really recommend the PC for people who wants install heavy games programs.	0.74
- The computer is nice, fast, light, ok.	0.12
- I bought the computer for my daughter.	0.05

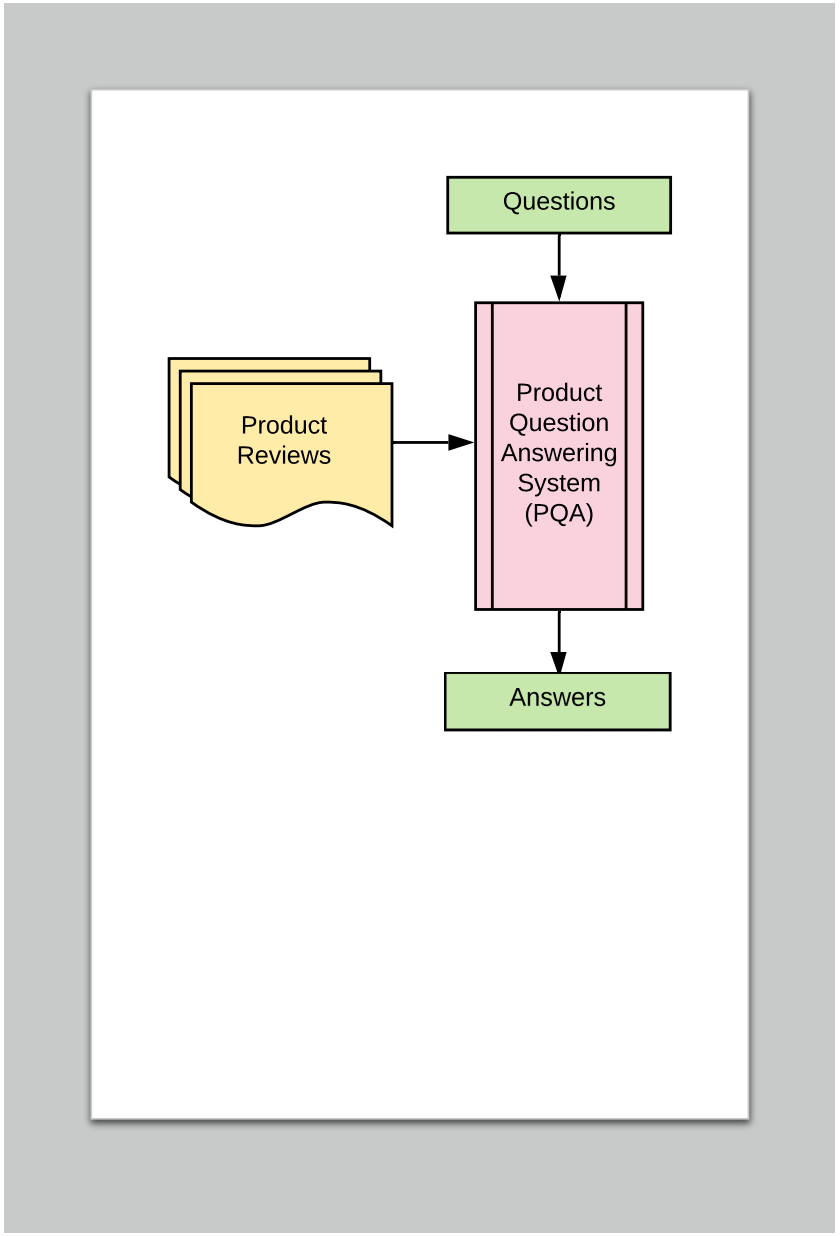
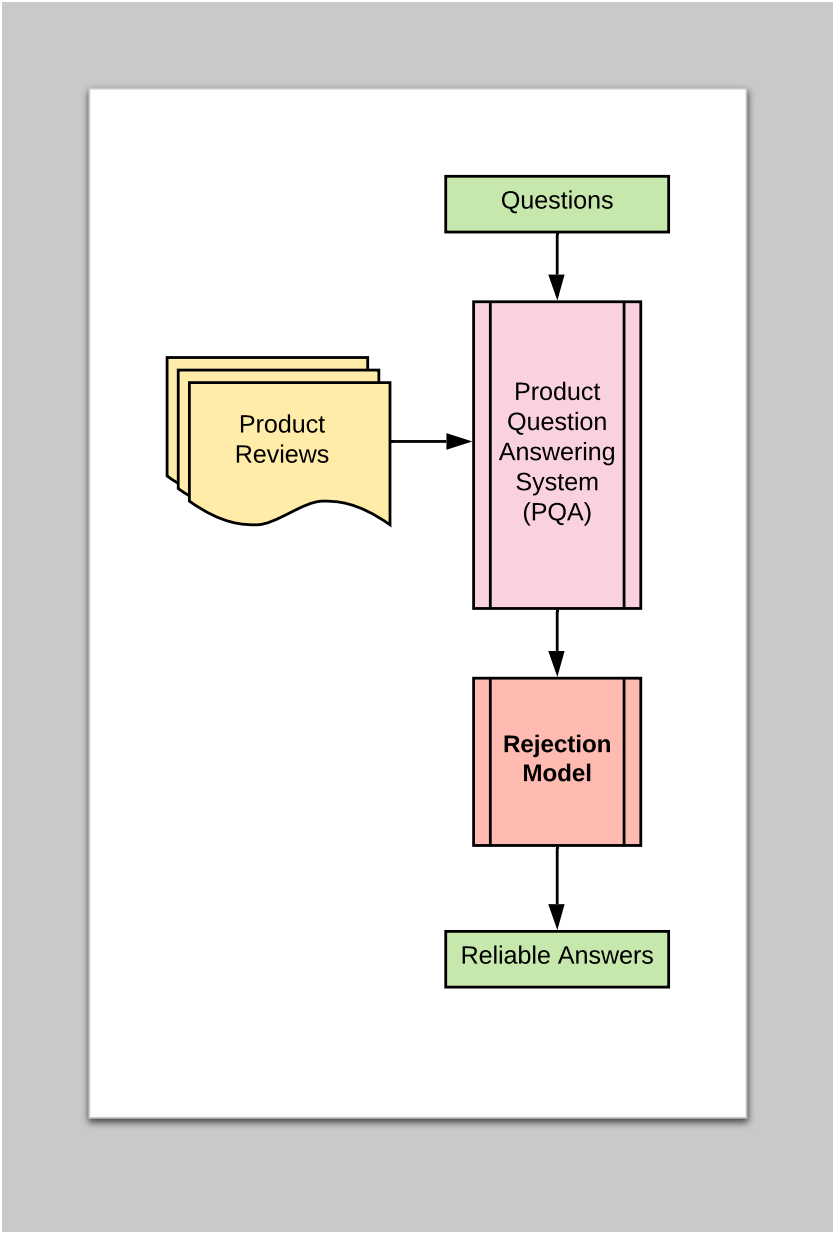


Table 1. Example of an answerable (Q1) and an unanswerable question (Q2). Green denotes high probability/confidence scores, and red otherwise.

Q1: What is the chain for on the side?			
Top 3 Ranked Reviews	Prob	Conf	Accept
 - This was driving me crazy but i see that another reviewer explained that grill has wire clip on chain to be used as extended match holder for igniting the gas if the spark mechanism fails to work or is worn out as sometimes happens with any gas grill.	0.99	0.82	✔
- PS Could not figure out the purpose of that little chain with the clip attached to the outside of the grill even after reading entire manual.	0.95	0.54	✘
- It is to replace an old portable that I have been using for about 10 years.'	0.91	0.40	✘
Q2: Does this Dell Inspiron 14R i14RMT-7475s come with dell's warranty?			
Top 3 Ranked Reviews	Prob	Conf	Accept
 - I don't really recommend the PC for people who wants install heavy games programs.	0.74	0.48	✘
- The computer is nice, fast, light, ok.	0.12	0.01	✘
- I bought the computer for my daughter.	0.05	0.00	✘

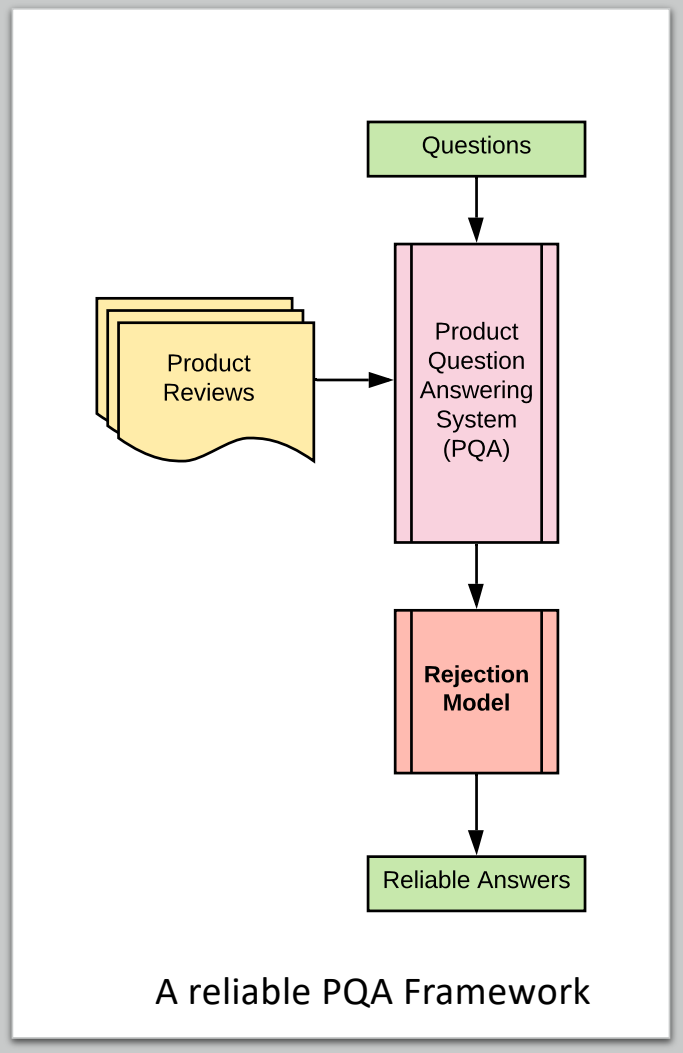


With a rejection model, a PQA model can reject unreliable reviews. By rejecting unreliable answers:

1. The returned results are more concise and accurate for **answerable** questions, e.g. [R,R **|** R,R,R].
2. Returning nil answers for **unanswerable** questions, e.g. [**|**R,R,R,R,R].

Key challenges:

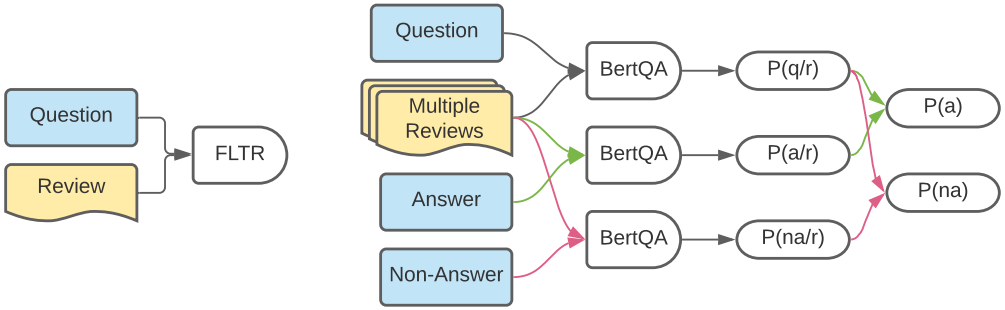
1. Choosing an appropriate model as the rejection model.
2. Identifying risks in PQA and choosing an appropriate risk function.
3. The lack of ground truth in PQA, i.e. there is limited data with annotated relevance scores between questions and reviews.

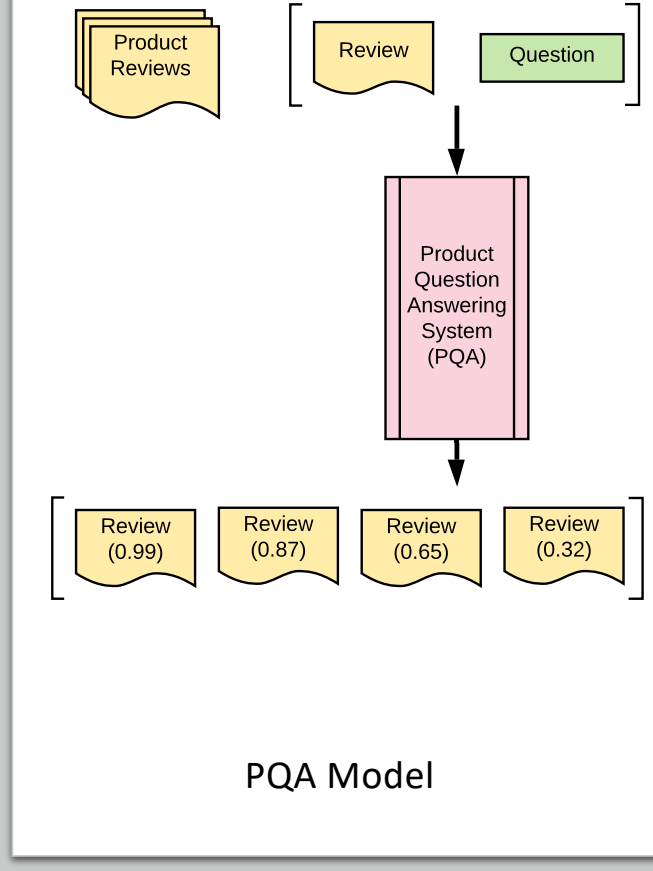


A reliable PQA Framework

PQA Models :

- **MOQA** [1]: a mixture of experts (MoE) based model (features like BM25+ , ROUGE-L and a learned bilinear scoring function).
- **FLTR** [2]: a BERT based classifier (zero-shot transfer).
- **BERTQA** [2]: an extension of MOQA, with BERT as the relevance function.





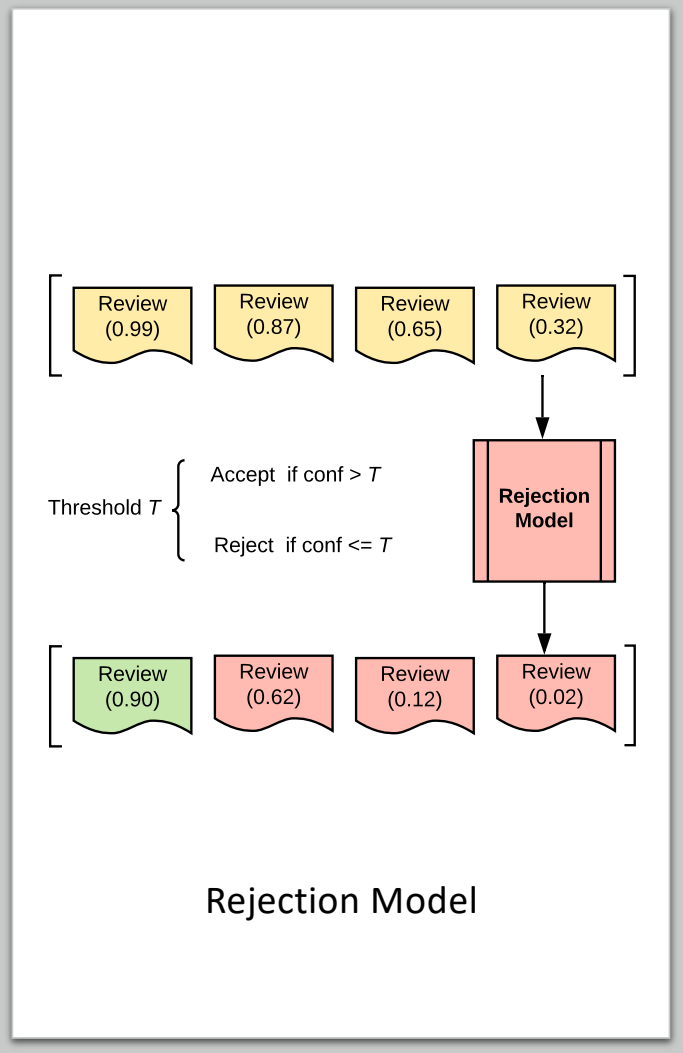
[1] McAuley, J. and Yang, A.: Addressing complex and subjective product-related queries with customer reviews. In: WWW (2016)
[2] Zhang, S., Lau, J.H., Zhang, X., Chan, J., Paris C.: Discovering Relevant Reviews for Answering Product-related Queries. In: ICDM (2019)

Rejection Model [1] is used to reduce risk (such as misclassification rate) by rejecting unconfident or unreliable predictions.

In a classification task, a model could make a wrong prediction for a difficult instance, particularly when the positive class probability is around 0.5 in a binary task. For medical applications such as tumour diagnostic, misclassification can have serious consequences. In these circumstances, rejection techniques are used to reduce misclassification rate by rejecting unconfident or unreliable predictions.

Rejection Model usually consists of two functions:

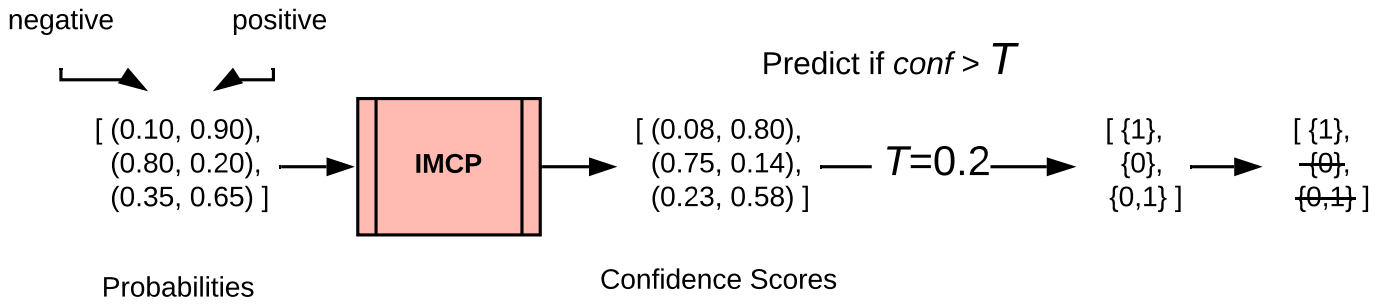
1. **Confidence function**: transforms probabilities into confidence scores.
2. **Risk function**: calculates the risk from wrong predictions and then tune a **threshold** *T* minimizing the risk on validation data.

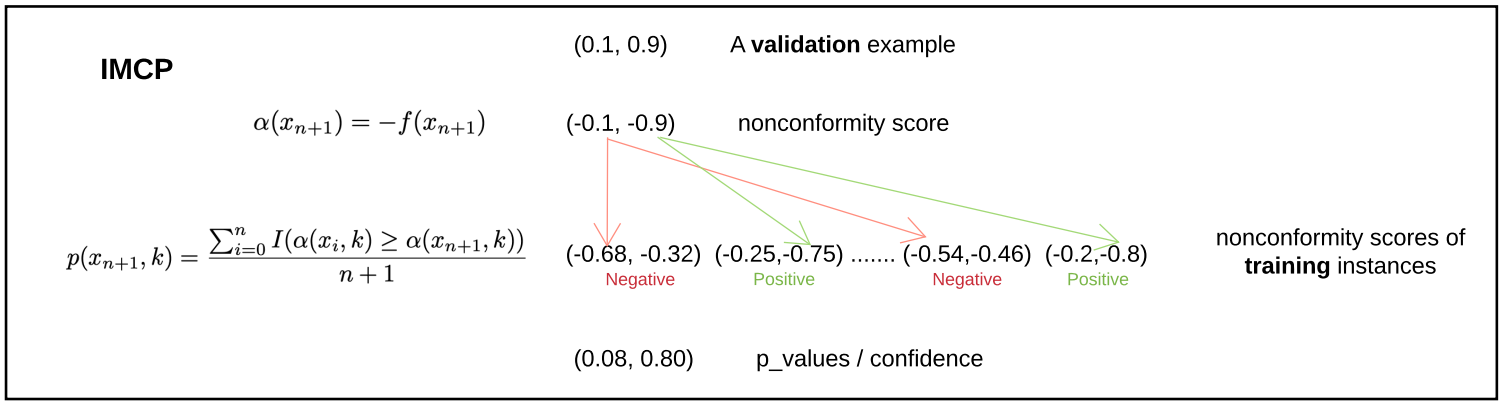


Rejection Model

1. Herbei, R. and Wegkamp, M.H.: Classification with reject option. The Canadian Journal of Statistics/La Revue Canadienne de Statistique. (2006)

Confidence function: Inductive Mondrian Conformal Predictor (**IMCP)** [2] is a modified Conformal Predictor [1] (a popular non-parametric model for reliable predictions), which is better at handling imbalanced data.





1. Vovk, V., Gammerman, A. and Shafer, G.: Algorithmic Learning in a Random World. Springer Science & Business Media. (2005)
2. Toccaceli, P. and Gammerman, A.: Combination of Inductive Mondrian Conformal Predictors. Machine Learning. (2019)

There are two forms of **risks** in PQA:

- Including irrelevant reviews as part of its returned results for an **answerable** question.
- Returning any reviews for an **unanswerable** question.

Q1: What is the chain for on the side?

Top 3 Ranked Reviews	Prob
- This was driving me crazy but i see that another reviewer explained that grill has wire clip on chain to be used as extended match holder for igniting the gas if the spark mechanism fails to work or is worn out as sometimes happens with any gas grill.	0.99
- PS Could not figure out the purpose of that little chain with the clip attached to the outside of the grill - even after reading entire manual.	0.95
- It is to replace an old portable that I have been using for about 10 years.'	0.91

Q2: Does this Dell Inspiron 14R i14RMT-7475s come with dell's wa

Top 3 Ranked Reviews	Prob
- I don't really recommend the PC for people who wants install heavy games programs.	0.74
- The computer is nice, fast, light, ok.	0.12
- I bought the computer for my daughter.	0.05

Risk function: NDCG' [1] $\in$ [0,1] is a variant of NDCG (normalized discounted cumulative gain). The key idea of NDCG' is to "quit while ahead": the returned document list should be truncated earlier rather than later, as documents further in the list are more likely to be irrelevant.

Table 4. NDCG' examples.

Question Type	Systems	Doc List	NDCG'
Answerable	System A	111	1.000
	System B	11100	0.971
	System C	11	0.922
Unanswerable	System A	$\emptyset$	1.000
	System B	00	0.500
	System C	000	0.431

1. System A > System B
2. System A, System B > System C

1. System A > System B > System C

Tuning *T* by maximizing NDCG'

[1] Liu, F., Moffat, A., Baldwin, T. and Zhang, X.: Quit while ahead: Evaluating truncated rankings. In: SIGIR. (2016)

We compare the following **methods:**

Without rejection:

- **Vanilla PQA Model**: top-10 reviews returned by a PQA model.

With rejection (**Our Solution**):

- **PQA Model + THRS**: tune a threshold based on the **probabilities** (top-10 reviews) returned by a PQA model.
- **PQA Model + IMCP**: tune a threshold based on the **confidences** (top-10 reviews) returned by IMCP.

PQA models: MOQA, FLTR and BERTQA

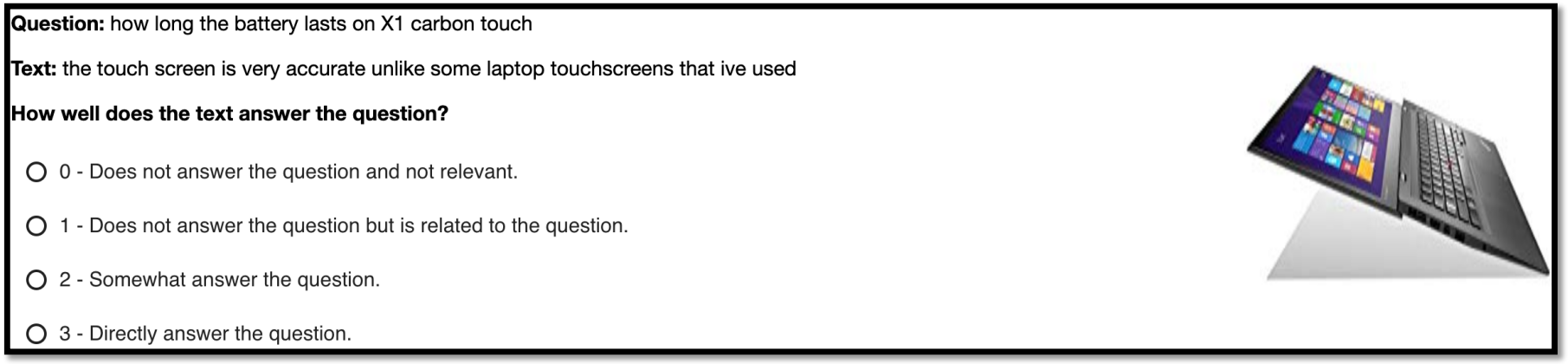
Data:

- Amazon QA dataset [1] for training PQA models.

- Annotated data (200 questions, 4,691 reviews) for evaluation. Each question/review pair is annotated by 3 workers.

Table 3. Answerable question statistics.

Relevance Threshold	2.00	2.25	2.50	2.75	3.00
#Relevant Reviews	640	351	175	71	71
#Answerable Questions	170	134	89	44	44
%Answerable Questions	85%	67%	45%	22%	22%



[1] McAuley, J. and Yang, A.: Addressing complex and subjective product-related queries with customer reviews. In: WWW (2016)

Relevance	Model	$\mathcal{N}_{A+U}$	$\mathcal{N}_A$	$\mathcal{N}_U$
≥ 2.00	MOQA	0.294	0.309	0.279
	MOQA+THRS	**0.319**	0.212	0.480
	MOQA+IMCP	0.318	0.212	0.477
	FLTR	0.372	0.495	0.279
	FLTR+THRS	**0.516**	0.400	0.666
	FLTR+IMCP	0.514	0.392	0.675
	BERTQA	0.360	0.464	0.279
	BERTQA+THRS	0.436	0.356	0.534
	BERTQA+IMCP	**0.447**	0.345	0.580
≥ 2.25	MOQA	0.264	0.249	0.279
	MOQA+THRS	**0.296**	0.179	0.489
	MOQA+IMCP	0.295	0.163	0.535
	FLTR	0.361	0.468	0.279
	FLTR+THRS	0.452	0.335	0.608
	FLTR+IMCP	**0.482**	0.329	0.705
	BERTQA	0.344	0.423	0.279
	BERTQA+THRS	0.373	0.293	0.477
	BERTQA+IMCP	**0.405**	0.310	0.530

Relevance	Model	$\mathcal{N}_{A+U}$	$\mathcal{N}_A$	$\mathcal{N}_U$
≥ 2.50	MOQA	0.243	0.211	0.279
	MOQA+THRS	**0.274**	0.165	0.453
	MOQA+IMCP	0.265	0.155	0.452
	FLTR	0.359	0.462	0.279
	FLTR+THRS	0.439	0.326	0.592
	FLTR+IMCP	**0.470**	0.316	0.699
	BERTQA	0.340	0.414	0.279
	BERTQA+THRS	**0.404**	0.308	0.530
	BERTQA+IMCP	0.387	0.294	0.510
≥ 2.75	MOQA	**0.235**	0.199	0.279
	MOQA+THRS	0.229	0.129	0.407
	MOQA+IMCP	0.213	0.107	0.423
	FLTR	0.333	0.397	0.279
	FLTR+THRS	0.409	0.272	0.615
	FLTR+IMCP	**0.416**	0.299	0.577
	BERTQA	0.330	0.390	0.279
	BERTQA+THRS	0.349	0.279	0.435
	BERTQA+IMCP	**0.388**	0.296	0.509

a) **With rejection** model (+THRS or + IMCP) is **better** than **without rejection** model.

b) For both FLTR and BERTQA, **+IMCP** is better than **+THRS** in most cases.

c) For MOQA, +IMCP seems worse than +THRS, it is may be due to MOQA producing an arbitrary (non-probabilistic) score for review.

Table 6. Reviews produced by FLTR, FLTR+THRS and FLTR+IMCP for an answerable (Q1) and unanswerable (Q2) question.

Q1: How long the battery lasts on X1 carbon touch?		
	Ground Truth	[3, 3]
	FLTR	[0, 0, 0, 0, 3, 0, 0, 0, 0, 0]
	FLTR+THRS	[0, 0, 0, 0, 3, 0, 0, 0]
	FLTR+IMCP	[0, 0, 0, 0, 3]
Q2: What type of memory SD card should I purchase to go with this?		
	Ground Truth	[ ]
	FLTR	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
	FLTR+THRS	[0, 0]
	FLTR+IMCP	[ ]

a) FLTR+IMCP rejects more irrelevant reviews and produces a concise list.

b) However, reject model does not modify the original ranking of the returned reviews. (Rejection model performs monotonic transformation.)

Conclusion

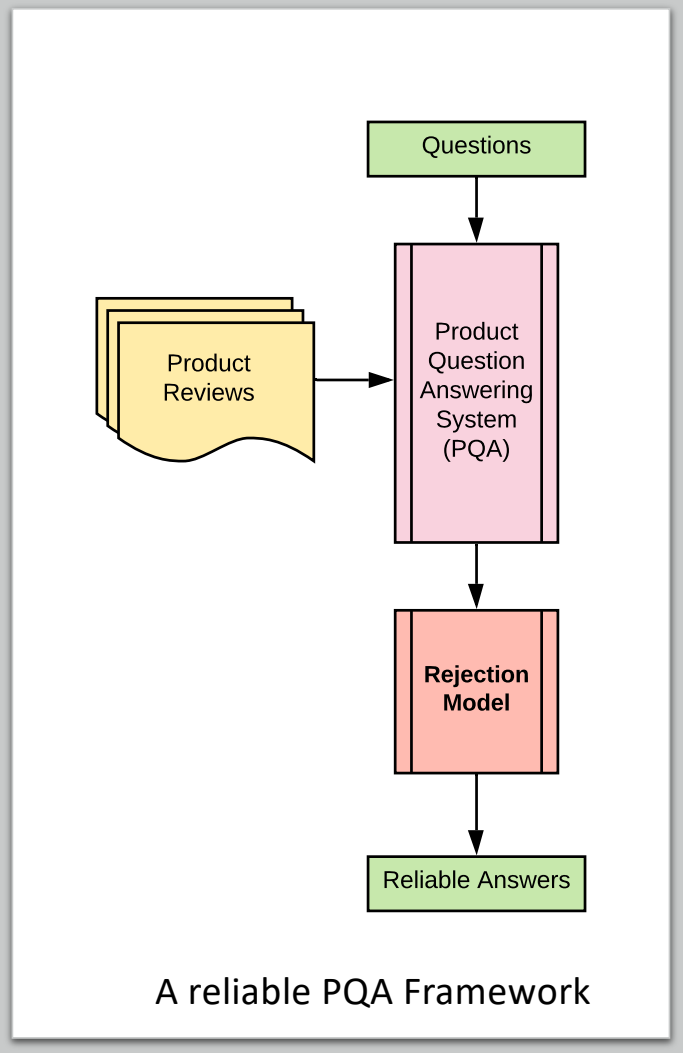
In this paper, we propose to incorporate conformal predictor as the rejection model to a PQA model to reject unreliable reviews. We found that using IMCP as the rejection model generally gives better results.

By rejecting unreliable answers:

1. The returned results are more concise and accurate for **answerable** questions.
2. Returning nil answers for **unanswerable** questions.

Code: https://github.com/zswvivi/icdm_pqa

Thanks and QA



A reliable PQA Framework