CliniQG4QA: Generating Diverse Questions for Domain Adaptation of Clinical Question Answering

Xiang Yue[1*], Xinliang Frederick Zhang[1,2*], Ziyu Yao[3], Simon Lin[4] and Huan Sun[1]

[1] The Ohio State University

[2] University of Michigan

[3] George Mason University

[4] Nationwide Children's Hospital

* These two authors contributed equally

Question Answering

What is COVID-19?

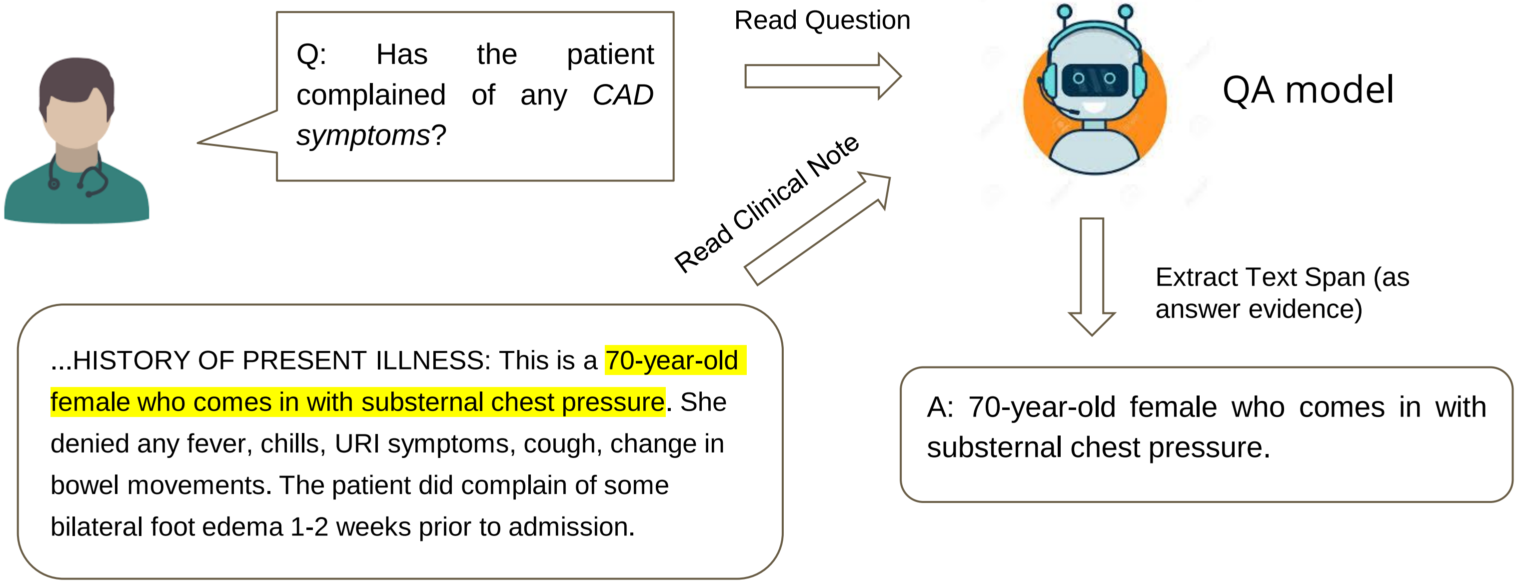
What is the risk of my child becoming sick with COVID-19?

What is the main cause of HIV-1 infection in children?

AUTOMATIC QUESTION ANSWERING

Clinical Reading Comprehension (CliniRC)

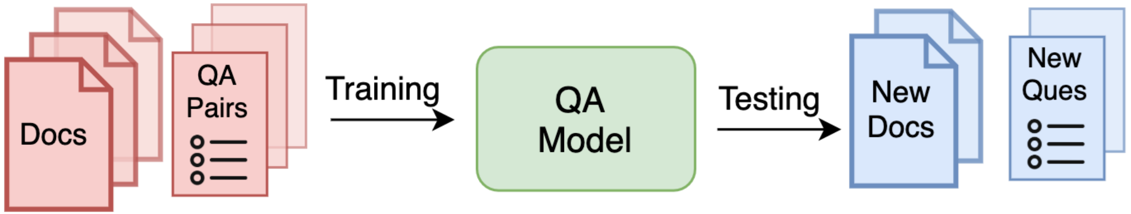
Automatically answer a user (e.g., doctor/clinician/researcher) question for a specific patient based on the patient clinical note.

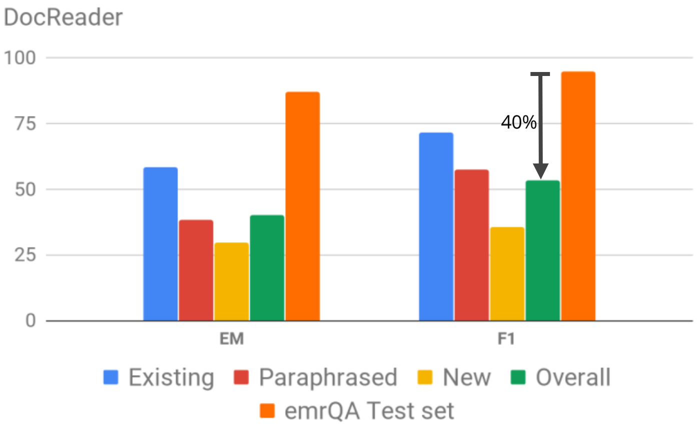


Generalization Issue

- A fully-trained QA model should **generalize** to a new environment

E.g.,





The model **struggles to generalize** to new questions and new contexts.

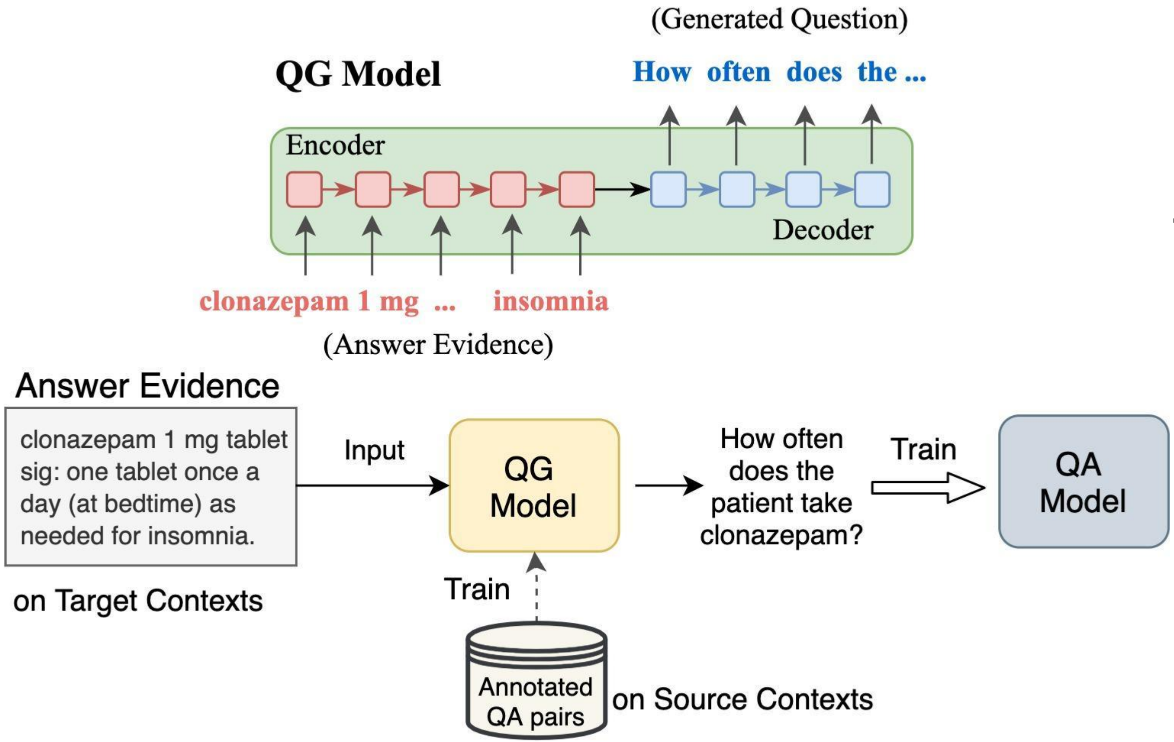
"Clinical Reading Comprehension: A Thorough Analysis of the emrQA Dataset", **Xiang Yue,** Bernal Jimenez Gutierrez and Huan Sun, **ACL 2020**

Crowdsourcing? Full Human Annotations?

Highly Impractical!

- Considerable medical expertise
- Data handling must be specifically designed
- Ethical issues
- Privacy concerns
- Time-consuming
- Costly
- …

Question Generation for Question Answering (QA)

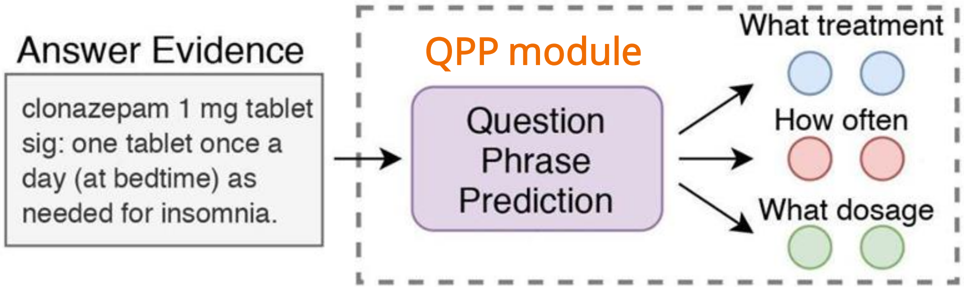


One issue we observe in our preliminary experiment:

The generated questions are less diverse, which are less useful for improving QA

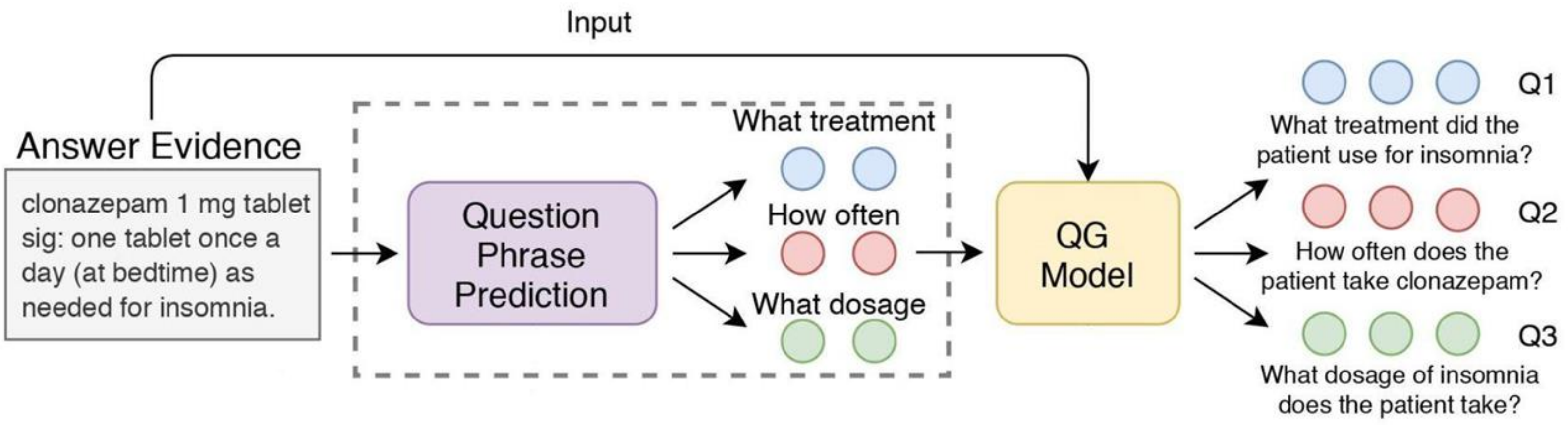
Our solution: Diverse Question Generation for QA

Step 1: Diverse question phrase generation via our QPP module



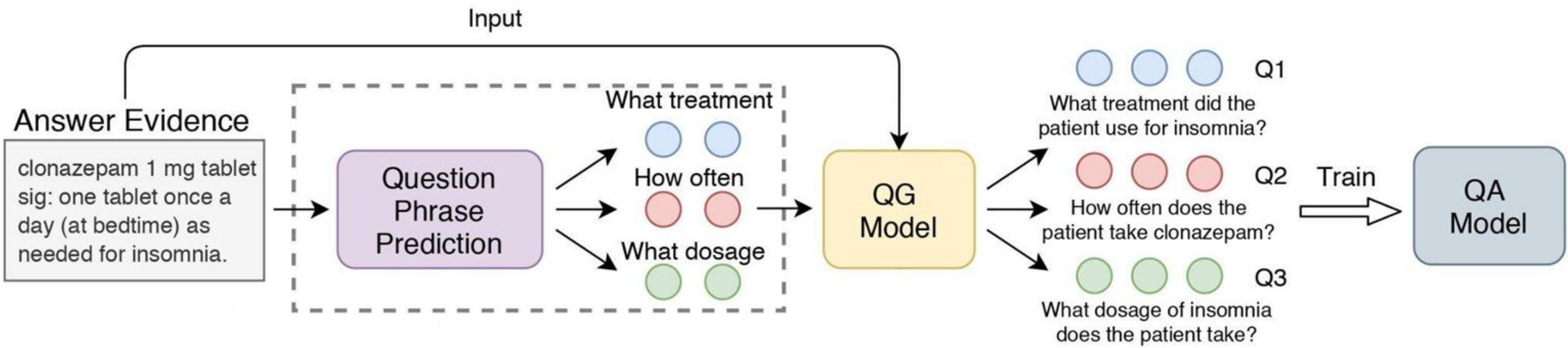
Our solution: Diverse Question Generation for QA

Step 2: QG model completes the rest of the question



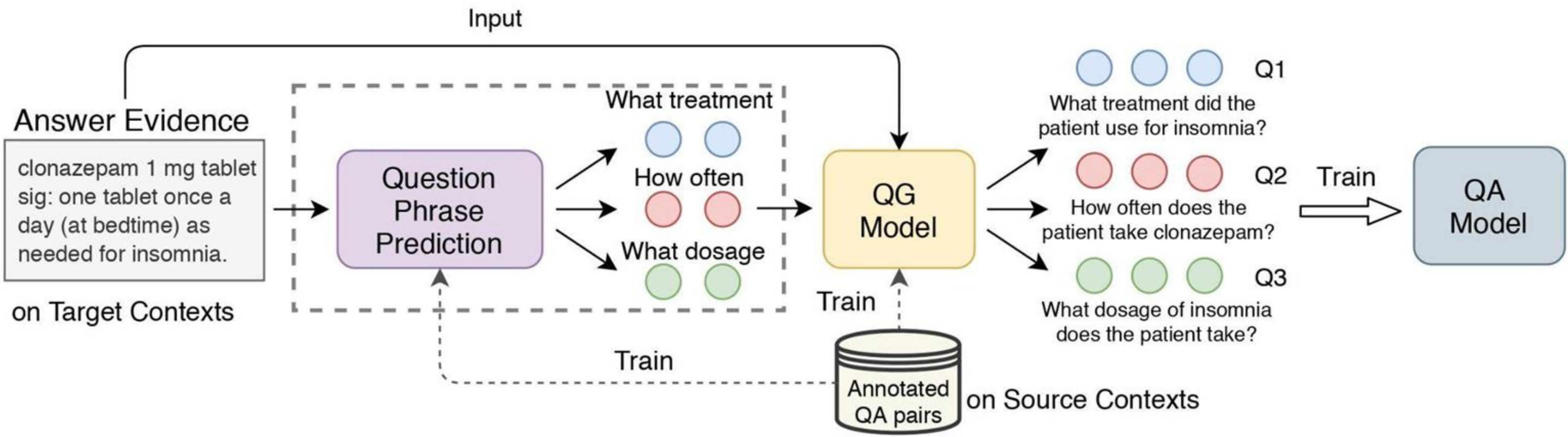
Our solution: Diverse Question Generation for QA

Step 3: Generated QA pairs are used to further train QA model



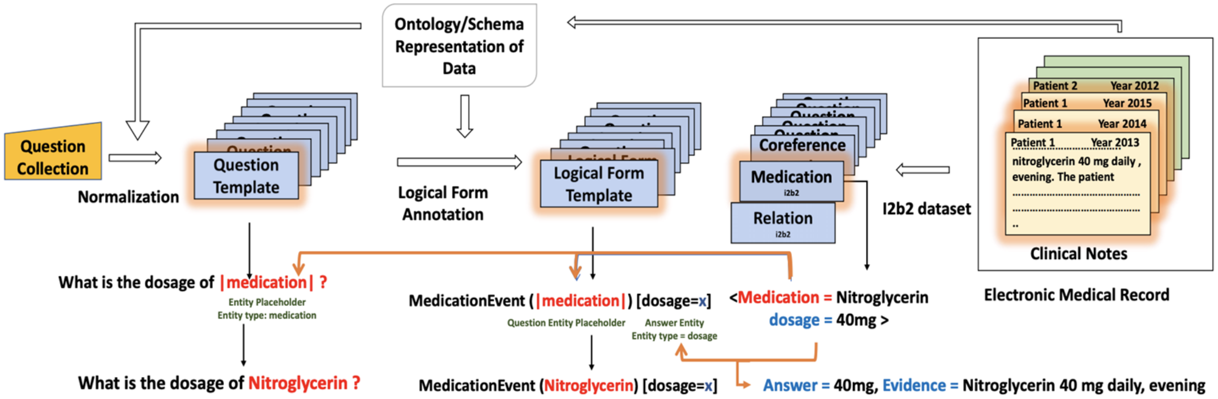
Our solution: Diverse Question Generation for QA

QG and QPP are trained on the source domain data



Datasets

Source domain data: *emrQA* [Pampari+ 2018]



Pampari, A., Raghavan, P., Liang, J., & Peng, J. (2018). emrqa: A large corpus for question answering on electronic medical records. EMNLP 2018

Datasets

Target domain Test data: We ask clinical experts to annotate *1,287 QA pairs* on ***MIMIC-III clinical texts*** for testing purpose.

- **Human-Generated (HG)** (*312 QA pairs*): Questions created by human experts.
- **Human-Verified (HV)** (*975 QA pairs*): Questions automatically generated by 3 base QG models and their variants used in this work, which are further verified by experts to ensure the correctness.

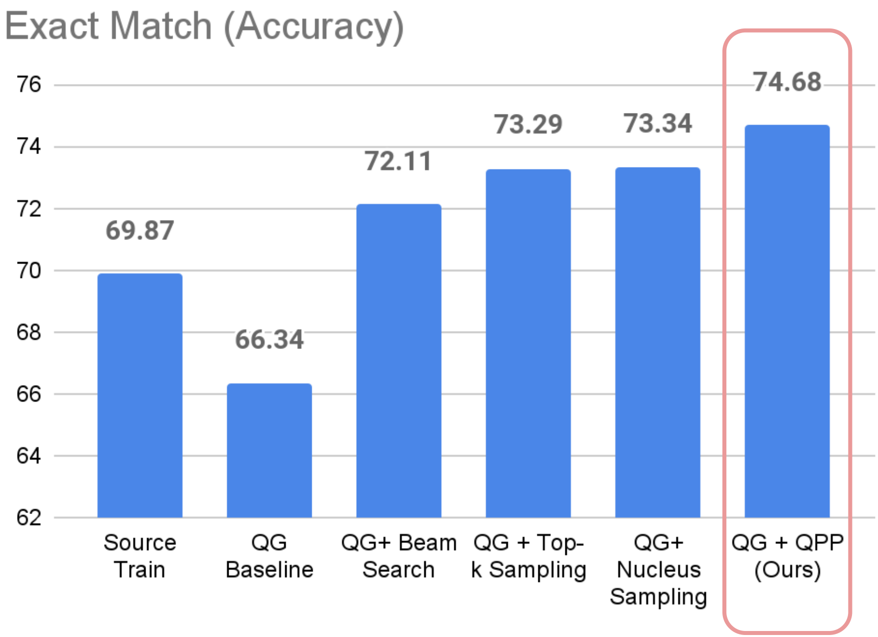
QA Results on New Documents

QA Datasets	DocReader [6]						ClinicalBERT [31]					
	Human Generated		Human Verified		Overall Test		Human Generated		Human Verified		Overall Test	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
emrQA [3]	69.87	83.66	61.44	78.82	63.48	79.99	69.23	82.83	61.23	78.56	63.17	79.59
NQG [10]	66.99	79.67	64.71	79.36	65.26	79.43	67.30	82.59	59.49	76.68	61.38	78.11
+ BeamSearch	71.15	83.07	67.07	81.21	68.07	81.66	68.91	84.26	63.17	79.17	64.56	80.40
+ Top-k Sampling	71.58	83.48	66.77	80.45	67.94	81.19	67.74	81.96	60.82	78.16	62.50	79.08
+ Nucleus Sampling	70.62	83.68	67.16	80.37	68.00	81.17	68.70	83.21	62.36	77.89	63.90	79.18
+ QPP (Ours)	**74.36**	**85.18**	**68.82**	**82.89**	**70.09**	**83.44**	**69.23**	**84.33**	**63.79**	**79.56**	**65.11**	**80.72**
NQG++ [15]	66.34	81.34	65.94	78.71	66.04	79.35	65.06	80.11	59.59	75.85	60.92	76.88
+ BeamSearch	72.11	84.56	68.10	80.09	69.07	81.17	68.26	83.70	64.61	80.30	65.50	81.12
+ Top-k Sampling	73.29	85.56	69.11	82.38	69.41	83.35	70.19	85.61	62.84	79.77	64.62	81.19
+ Nucleus Sampling	73.34	84.95	68.94	81.72	70.01	82.51	70.19	84.72	63.93	79.54	65.45	80.80
+ QPP (Ours)	**74.68**	**85.92**	**70.05**	**83.47**	**71.10**	**84.06**	**70.83**	**85.76**	**65.33**	**80.64**	**66.67**	**81.88**
BERT-SQG [34]	70.19	81.47	66.05	79.64	67.05	80.08	65.06	82.20	59.59	78.04	60.92	79.05
+ BeamSearch	73.71	84.44	68.71	81.98	69.93	82.58	67.31	82.54	61.94	79.02	63.25	79.88
+ Top-k Sampling	72.81	84.16	69.20	82.24	70.07	82.71	69.12	84.20	60.44	78.27	62.55	79.71
+ Nucleus Sampling	70.73	83.60	68.56	81.80	69.09	82.24	67.74	83.16	61.61	78.74	63.09	79.81
+ QPP (Ours)	**74.36**	**85.53**	**70.77**	**83.60**	**71.64**	**84.07**	**69.23**	**85.38**	**64.21**	**80.53**	**65.43**	**81.71**

QA Results on New Documents

QA Datasets	DocReader [6]						ClinicalBERT [31]					
	Human Generated		Human Verified		Overall Test		Human Generated		Human Verified		Overall Test	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
emrQA [3]	69.87	83.66	61.44	78.82	63.48	79.99	69.23	82.83	61.23	78.56	63.17	79.59
NQG [10]	66.99	79.67	64.71	79.36	65.26	79.43	67.30	82.59	59.49	76.68	61.38	78.11
+ BeamSearch	71.15	83.07	67.07	81.21	68.07	81.66	68.91	84.26	63.17	79.17	64.56	80.40
+ Top-k Sampling	71.58	83.48	66.77	80.45	67.94	81.19	67.74	81.96	60.82	78.16	62.50	79.08
+ Nucleus Sampling	70.62	83.68	67.16	80.37	68.00	81.17	68.70	83.21	62.36	77.89	63.90	79.18
+ QPP (Ours)	**74.36**	**85.18**	**68.82**	**82.89**	**70.09**	**83.44**	**69.23**	**84.33**	**63.79**	**79.56**	**65.11**	**80.72**
NQG++ [15]	66.34	81.34	65.94	78.71	66.04	79.35	65.06	80.11	59.59	75.85	60.92	76.88
+ BeamSearch	72.11	84.56	68.10	80.09	69.07	81.17	68.26	83.70	64.61	80.30	65.50	81.12
+ Top-k Sampling	73.29	85.56	69.11	82.38	69.41	83.35	70.19	85.61	62.84	79.77	64.62	81.19
+ Nucleus Sampling	73.34	84.95	68.94	81.72	70.01	82.51	70.19	84.72	63.93	79.54	65.45	80.80
+ QPP (Ours)	**74.68**	**85.92**	**70.05**	**83.47**	**71.10**	**84.06**	**70.83**	**85.76**	**65.33**	**80.64**	**66.67**	**81.88**
BERT-SQG [34]	70.19	81.47	66.05	79.64	67.05	80.08	65.06	82.20	59.59	78.04	60.92	79.05
+ BeamSearch	73.71	84.44	68.71	81.98	69.93	82.58	67.31	82.54	61.94	79.02	63.25	79.88
+ Top-k Sampling	72.81	84.16	69.20	82.24	70.07	82.71	69.12	84.20	60.44	78.27	62.55	79.71
+ Nucleus Sampling	70.73	83.60	68.56	81.80	69.09	82.24	67.74	83.16	61.61	78.74	63.09	79.81
+ QPP (Ours)	**74.36**	**85.53**	**70.77**	**83.60**	**71.64**	**84.07**	**69.23**	**85.38**	**64.21**	**80.53**	**65.43**	**81.71**

QA Results on New Documents



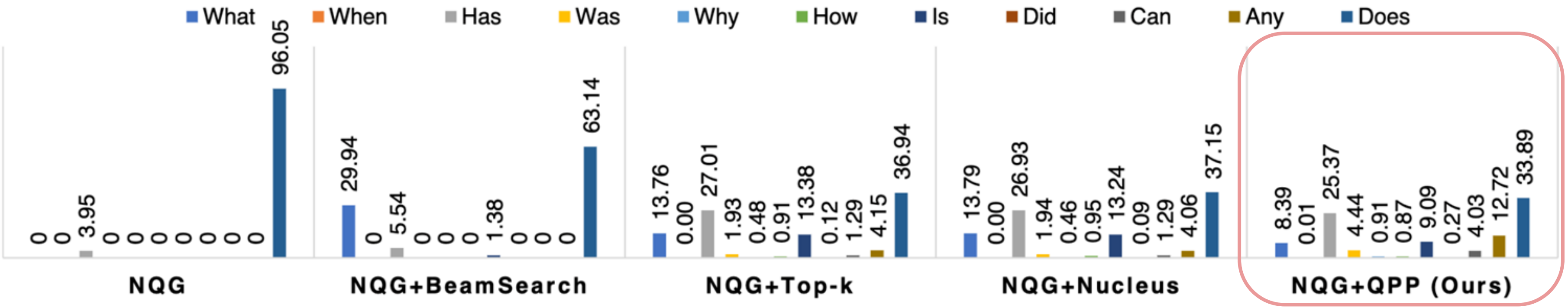
QG baseline: NQG++
QA baseline: DocReader

Dataset: Human-Generated (HG) Set

Our diverse QG (QG+QPP) method:

1) 5% absolute improvement over source training
2) 1%-2% absolute gain over other diverse generation strategies

Why QG Boosts QA on New Documents?



Diverse Questions Matter!

Why QG Boosts QA on New Documents?

QA Example from MIMIC-III
Context: ... he was guaiac negative on admission. hematocrit remained stable overnight. **5. abd pain: suspect secondary to chronic pancreatitis.** amylase unchanged from previous levels. ...
Question: Why did the patient get abd pain?
Answer by QA model trained on
-emrQA: 5. abd pain
-NQG: 5. abd pain:
-NQG+BeamSearch: 5. abd pain:
-NQG+Top-k: 5. abd pain:
-NQG+Nucleus: 5. abd pain:
-NQG+QPP: 5. abd pain: suspect secondary to chronic pancreatitis.

QG Example from MIMIC-III
Context: ... the patient was taking at home prior to admission were not restarted. **25. acetaminophen 325-650 mg po/ng q6h:prn pain** 26. dabigatran etexilate 150 mg po bid...
Questions generated by
-NQG: **Does** the patient have any pain?
-NQG+BeamSearch: **Does** the patient have any pain history? **Does** the patient have pain? **Does** the patient have any pain?
-NQG+Top-k: **Has** the patient ever had any pain? **Has** the patient ever reported pain? **Does** the patient have a history pain?
-NQG+Nucleus: **Has** the patient ever gone into pain? **What** happened when she was given morphine? Is there mention pain anywhere in the record?
-NQG+QPP: **Why** did the patient have acetaminophen? **What** treatment has the patient had for his pain? **How** was pain treated? **Does** the patient have any pain? ...

Diverse Questions Matter!

Summary

- Generating diverse QA pairs on the target contexts
  - QPP module plays an important role in generating *diverse* questions
  - *Diverse* synthetic data improves clinical QA performance on the target clinical documents
- **Future Work**
  - Test our method on more target datasets
  - Explore more advanced QG methods

- Code is available at:

  https://github.com/sunlab-osu/CliniQG4QA/

- Data is available at:

  https://physionet.org/content/mimic-iii-question-answer/1.0.0/

Git Repo QR code

Data QR code

Thanks!

Questions?

Research Supported by NSF, PCORI, ARO, Ohio Supercomputer Center