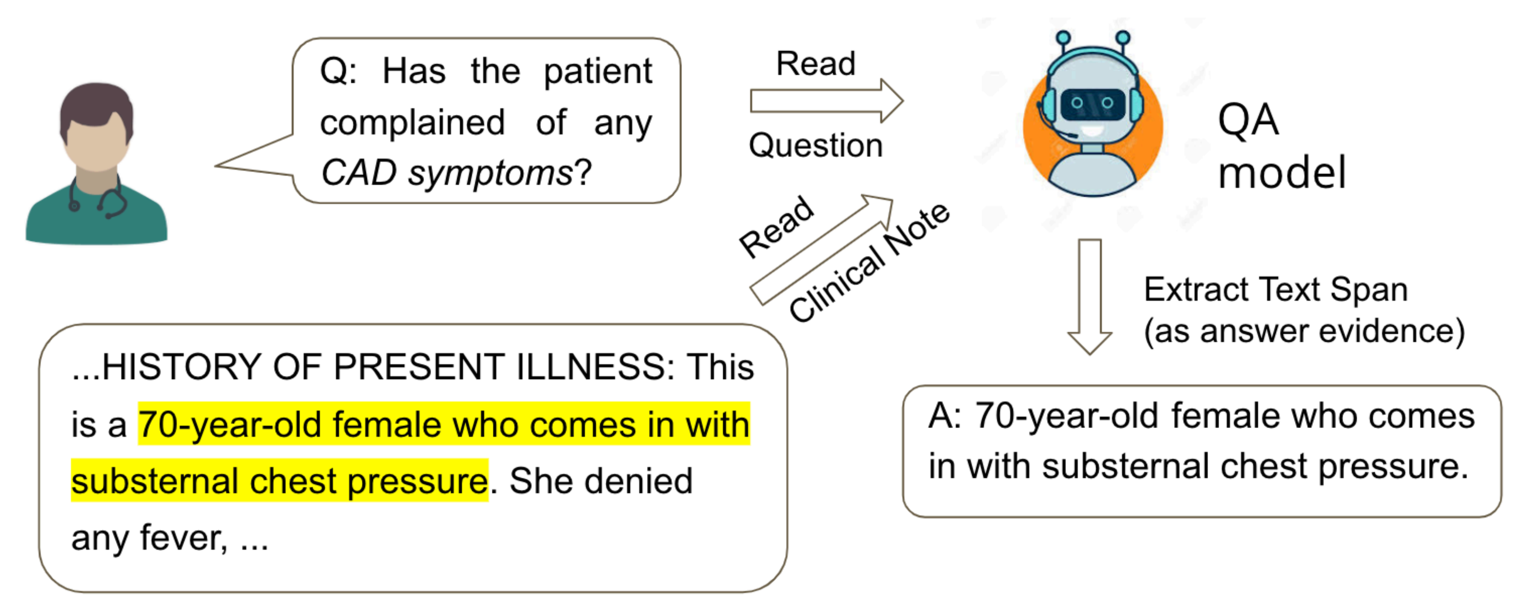
CliniQG4QA: Generating Diverse Questions for Domain Adaptation of Clinical Question Answering

Xiang Yue[1*], Xinliang (Frederick) Zhang[1*], Ziyu Yao[1], Simon Lin[2] and Huan Sun[1]
[1]The Ohio State University, [2]Nationwide Children's Hospital
[*]The first two authors contributed equally

Introduction

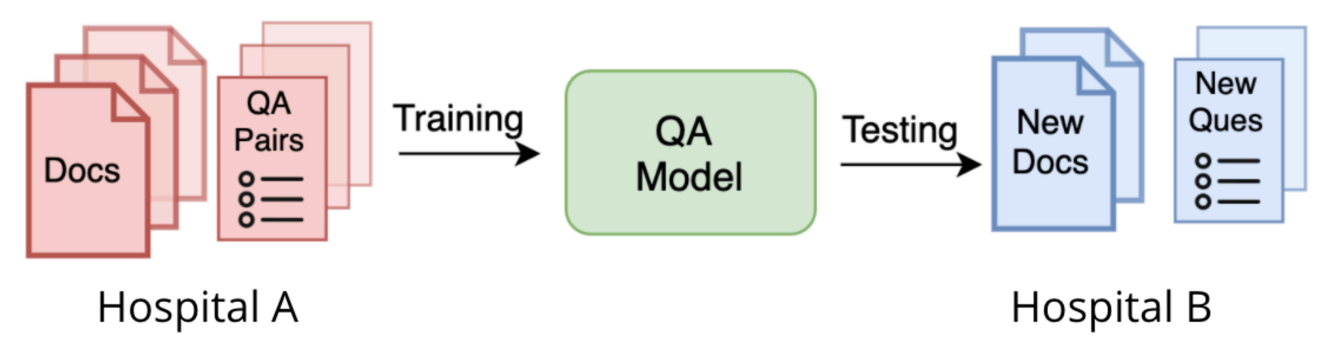
Clinical Question Answering (Reading Comprehension)

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

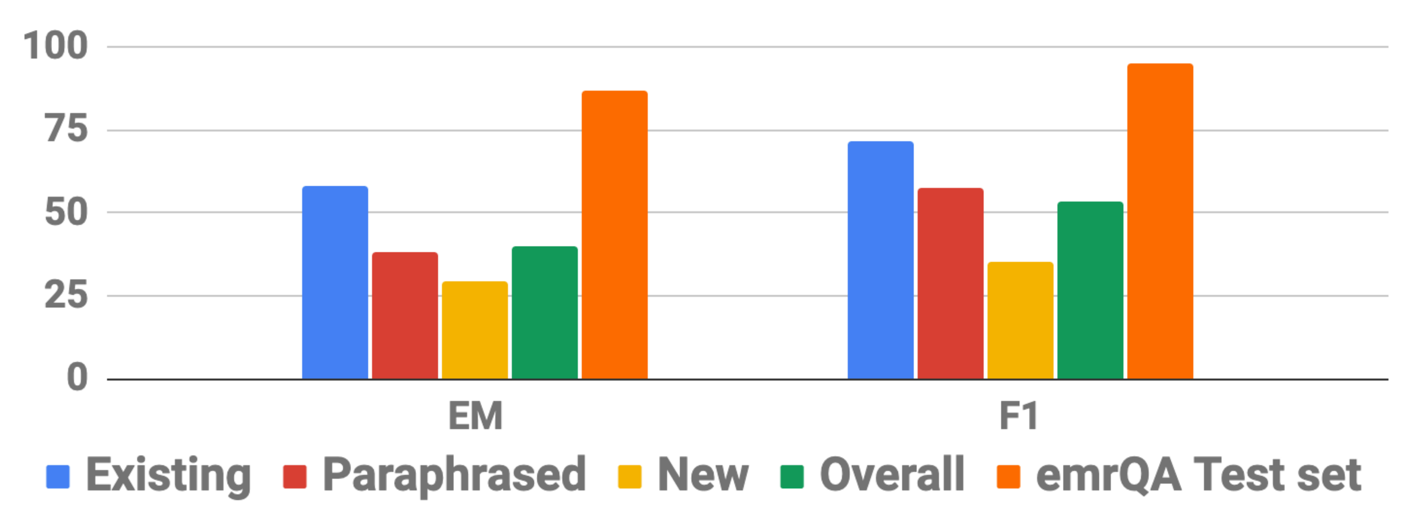


Generalization Issue

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



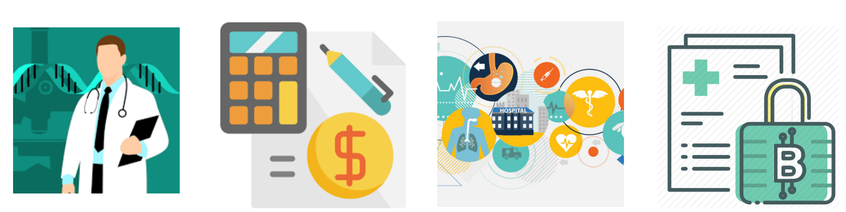
- However, according to (Yue et al., 2020), a base QA model trained on the emrQA dataset struggles to answer questions on the MIMIC-III dataset (40% drop overall compared with the original emrQA test set)



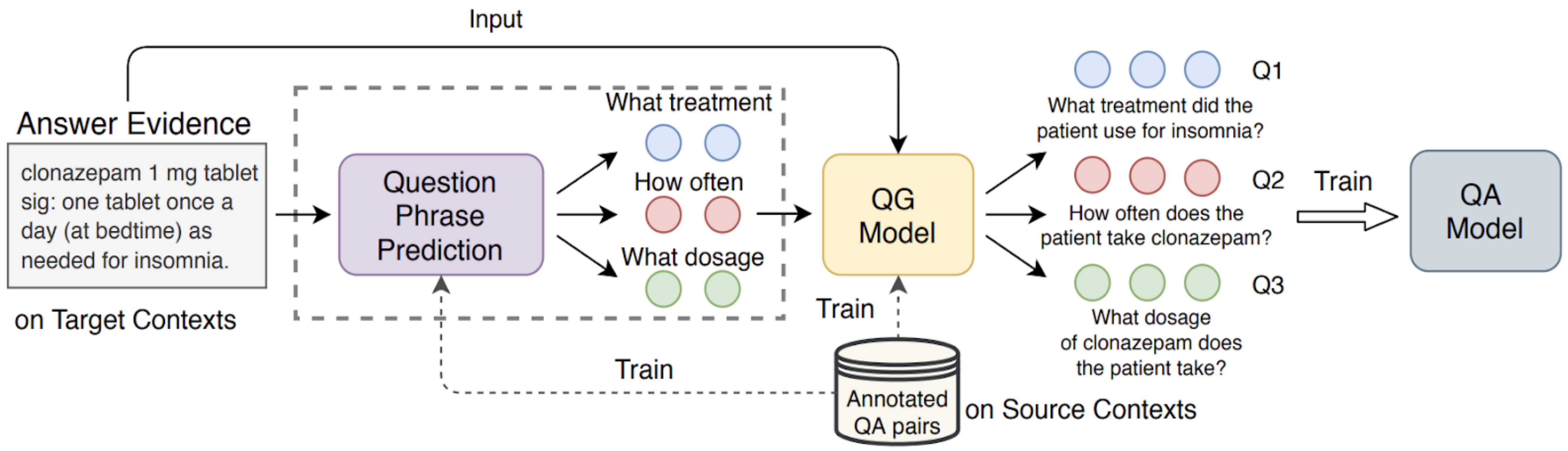
Annotations in the clinical domain

- Medical expertise
- Ethical issues
- Privacy concerns
- Time-consuming
- Costly

Highly Impractical!



Method



Algorithm 1 CliniQG4QA training procedure
Pretraining Stage
1: Train *Answer Evidence Extractor (AEE)* based on the *source* data
2: Obtain question phrases from source questions and train *Question Phrase Prediction (QPP)* module on the *source* data
3: Train a *QPP-enhanced QG* model on the *source* data
Training Stage
4: Use *AEE* to extract potential answer evidences on the *target* context
5: Use *QPP* to predict potential question phrases set
6: Use *QPP-enhanced QG* to generate diverse questions
7: Train a *QA* model on synthetic *target* data

Datasets

(Question / Context)	emrQA	MIMIC-III
# Train	781,857 / 337	- / 337
# Dev	86,663 / 41	8,824 / 40
# Test	98,994 / 42	1,287 / 36
# Total	967,514 / 420	- / 413
for purpose of	QG & QA (source)	QA (target)

- **MIMIC-III Test set Annotation**

Machine-generated QA pairs by 9 QG models are provided to experts as references.
Human-generated Questions: they are highly encouraged to create new questions.
Human-verified Questions: if they do find the machine-generated questions make sense they can keep them.

Conclusions

- CliniQG4QA leverages QG to synthesize QA pairs on new clinical contexts and boosts QA models without requiring manual annotations.
- Our question phrase prediction (QPP) module can be used together with most existing QG models to diversify their generation.
- QA corpus generated by our framework is helpful and that the QPP module plays a crucial role in achieving the gain.

Results

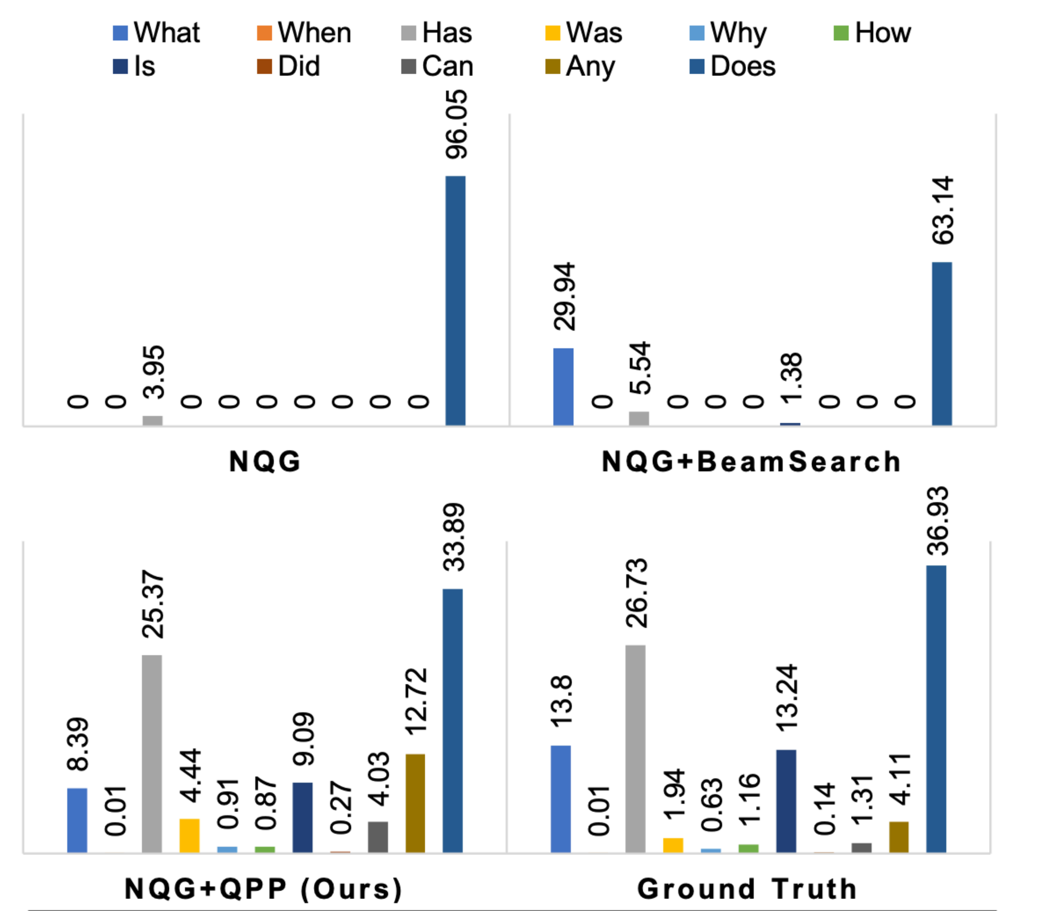
- **Can Generated Questions Help QA on New Contexts?**

QA Datasets	DocReader Chen et al. (2017)					
	Human Verified		Human Generated		Overall Test	
	EM	F1	EM	F1	EM	F1
emrQA (Pampari 2018)	61.44	78.82	69.87	83.66	63.48	79.99
NQG (Du 2017)	64.71	79.36	66.99	79.67	65.26	79.43
+ BeamSearch	67.07	81.21	71.15	83.07	68.07	81.66
+ QPP (Ours)	**68.82**	**82.89**	**74.68**	**85.18**	**70.09**	**83.44**
NQG++ (Zhou 2017)	65.94	78.71	66.34	81.34	66.04	79.35
+ BeamSearch	68.10	80.09	72.11	84.56	69.07	81.17
+ QPP (Ours)	**70.05**	**83.47**	**74.36**	**85.92**	**71.10**	**84.06**
BERT-SQG (Chan 2019)	66.05	79.64	70.19	81.47	67.05	80.08
+ BeamSearch	68.71	81.98	73.71	84.44	69.93	82.58
+ QPP (Ours)	**70.77**	**83.60**	**74.36**	**85.53**	**71.64**	**84.07**

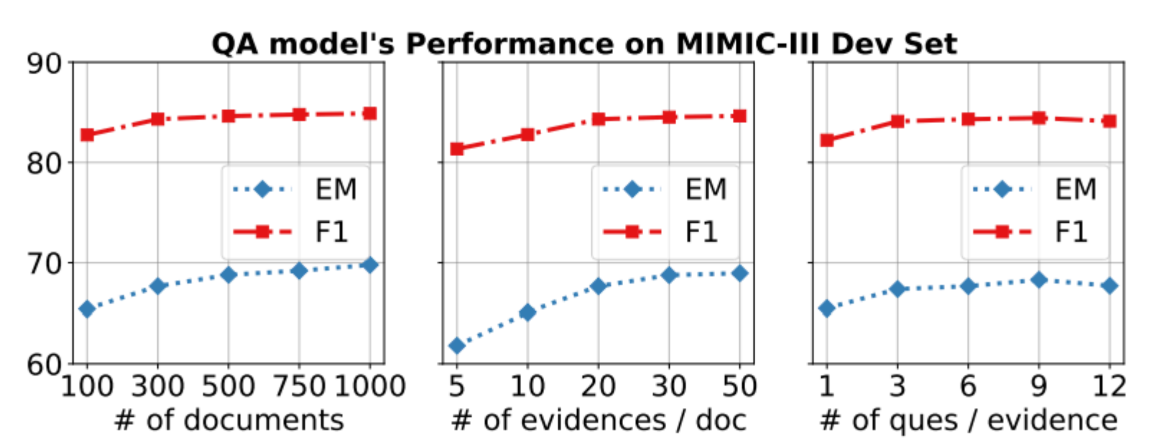
- **Diverse Questions Really Matter: Two Real Cases**

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

- **A Closer Look at Generated Question Types**



- **Why QG Boosts QA on New Contexts?**



References

[1] Yue, Xiang, et al.,. "Clinical Reading Comprehension: A Thorough Analysis of the emrQA Dataset." ACL 2020.
[2] Pampari, Anusri, et al. "emrqa: A large corpus for question answering on electronic medical records." EMNLP 2018.
[3] Du, Xinya, Junru Shao, and Claire Cardie. "Learning to ask: Neural question generation for reading comprehension." ACL 2017.
[4] Zhou, Qingyu, et al. "Neural question generation from text: A preliminary study." National CCF Conference on Natural Language Processing and Chinese Computing. Springer, Cham, 2017.
[5] Chan, Ying-Hong, and Yao-Chung Fan. "A Recurrent BERT-based Model for Question Generation." MRQA Workshop. 2019.
[6] Chen, Danqi, et al. "Reading wikipedia to answer open-domain questions." ACL 2017
[7] Johnson, Alistair EW, et al. "MIMIC-III, a freely accessible critical care database." Scientific data 3.1 (2016): 1-9.

Acknowledgement

pcori
NSF
Ohio Supercomputer Center

Contact & Code

- Contact:
yue.149@osu.edu, zhang.9975@osu.edu
sun.397@osu.edu
- Code:
https://github.com/sunlab-osu/CliniQG4QA