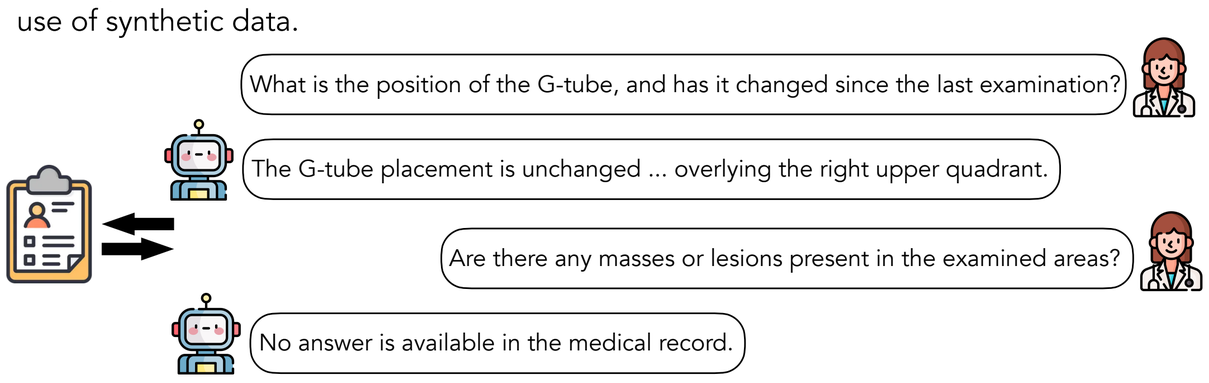
Give me Some Hard Questions: Synthetic Data Generation for Clinical QA

Fan Bai[J], Keith Harrigian[J], Joel Stremmel[O], Hamid Hassanzadeh[O], Ardavan Saeedi[O], Mark Dredze[J]

Motivation

Background: Clinical QA systems provide doctors better access to EHRs, but their development relies on **human-annotated QA pairs with privacy risks**, incentivizing the use of synthetic data.



Problem: Naive prompting yields **simple questions that closely mirror the input context**, failing to capture the complexity of clinical scenarios.

Generate 10 questions from a medical professional's viewpoint for the radiology report below.

FINDINGS: The **G-tube** placement is unchanged compared to the prior study with the tip of the pigtail catheter overlying the right upper quadrant. **Subcutaneous air** is still present. This was a supine film, and therefore, there is limited assessment for **free air**. Motion somewhat limits the evaluation, but the ...

1. What is the position of the **G-tube**, and has it changed since the last examination?
2. Is there any evidence of **free air** in the abdominal cavity, indicating a possible perforation?
3. Is the **subcutaneous air** present in the current film, and has it increased or decreased since the last study?
4. ...

Contributions

Contribution Summary:

- Propose a two-step method to **generate challenging questions** for Clinical QA: 1) **summarize the input using a predefined schema**; 2) instruct LLMs to generate **questions that do not overlap with the summary**
- Demonstrate the effectiveness of our methods, on **two Clinical QA datasets, RadQA and MIMIC-QA** and verify that our **generated questions are more challenging**.

Questions Generated by Our Method:

1. Is there any evidence of **gastrointestinal perforation**?
2. Are there any signs of infection or inflammation in the abdominal region?
3. Can we identify any abnormalities in the **lung fields** that might explain the patient's symptoms?
4. Is there any indication of **fluid accumulation** in the abdominal cavity?
5. Are there any masses or **lesions** present in the examined areas?
6. ...

Our method generates more challenging questions (**introducing relevant but unseen concepts, highlighted in red**).

Paper:
https://arxiv.org/pdf/2412.04573

Code & Data:
https://github.com/bflashcp3f/synthetic-clinical-qa

Paper | Code

Method

1. Input Summarization

Input

EHR Document

<radiology_report>
FINDINGS: The G-tube placement is unchanged compared to the prior study with .... Subcutaneous air is still present. This was a supine film, and therefore, there is limited assessment for free air. Motion somewhat limits the evaluation, but the bowel gas pattern appears grossly
</radiology_report>

Summarization Schema

Please generate a summary for the radiology report above to cover 5 following aspects: **"symptoms", "medical_conditions", "areas_examined", "patient_medical_history",** and **"diagnostic_techniques"**, and output in a JSON format.

Output

Attributed Summary

```
{
    "symptoms": ["abdominal pain"],
    "medical_conditions": ["stroke", "volume loss"],
    "areas_examined": ["right upper quadrant",
    "colon"],
    "patient_medical_history": ["stroke", "G-tube
    placement"],
    "diagnostic_techniques": ["supine film"],
}
```

2. Question Generation

Input

Attributed Summary

```
{
    "symptoms": ["abdominal pain"],
    "medical_conditions": ["stroke", "volume loss"],
    "areas_examined": ["right upper quadrant",
    "colon"],
    "patient_medical_history": ["stroke", "G-tube
    placement"],
    "diagnostic_techniques": ["supine film"],
}
```

Quest. Gen. Prompt

Generate 10 questions ... . **Make sure that the generated questions do not contain any words from the report** ...

Output

Generated Questions

1. Is there any evidence of gastrointestinal perforation?
2. Are there any signs of infection or inflammation in the abdominal region?
3. Can we identify any abnormalities in the lung fields that might explain the patient's symptoms?
4. Is there any indication of fluid accumulation in the abdominal cavity?
5. Are there any masses or lesions present in the examined areas?
6. ...

Experiments

Datasets (MIMIC-III database)

- RadQA (Soni et al., 2022): 1009 radiology reports, 3074 questions
- MIMIC-QA (Yue et al., 2021): 36 clinical notes, 1287 questions

Models

- Data-generation LLMs**:** gpt-4o and Llama3-8B-Instruct
- Fine-tuning Clinical QA model: BioClinRoBERTa-large

Metrics

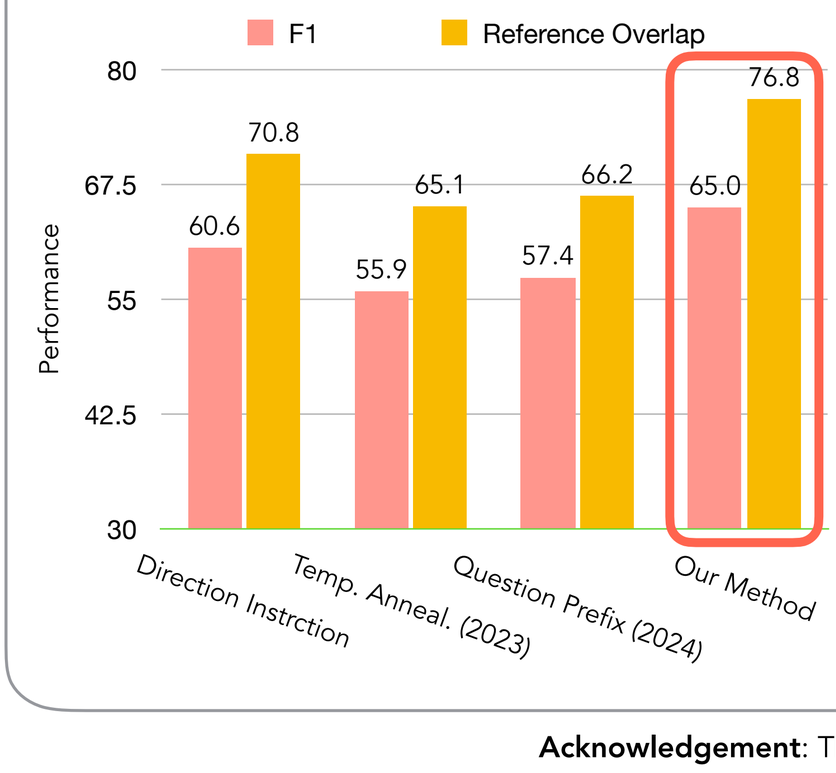
- Standard: Exact Match and F1
- New: Reference Overlap (overlap with the gold answer in the character index)

Baselines

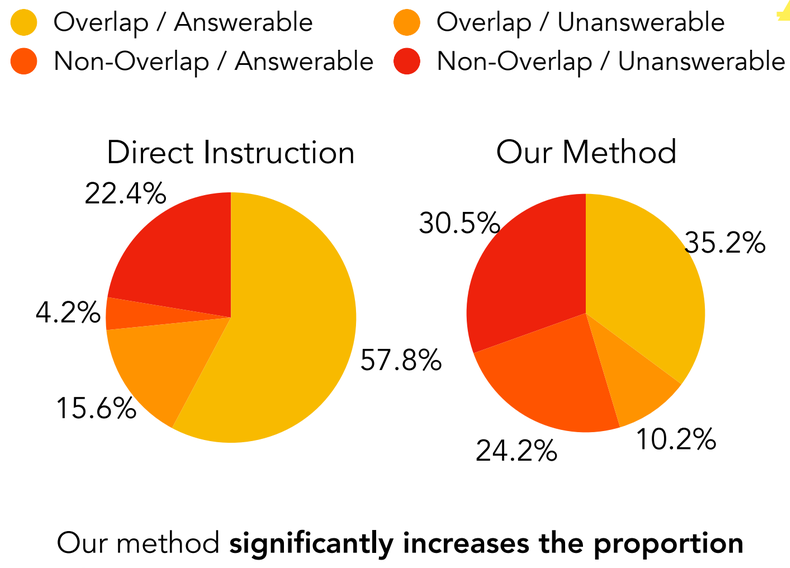
- Direct Instruction: instruct LLM to generate questions directly from the input, e.g., *"generate 10 questions from a medical professional's viewpoint."*
- Temperature Annealing: anneal the temperature of nucleus sampling from 0 to 1 during question generation to promote diversity.
- Question Prefix: generate questions with different prefixes, e.g., *'Please ensure each question starts with a different prefix, such as "is," "does," "has," ...'*

Results & Analyses

RadQA Results (data generated by gpt-4o)

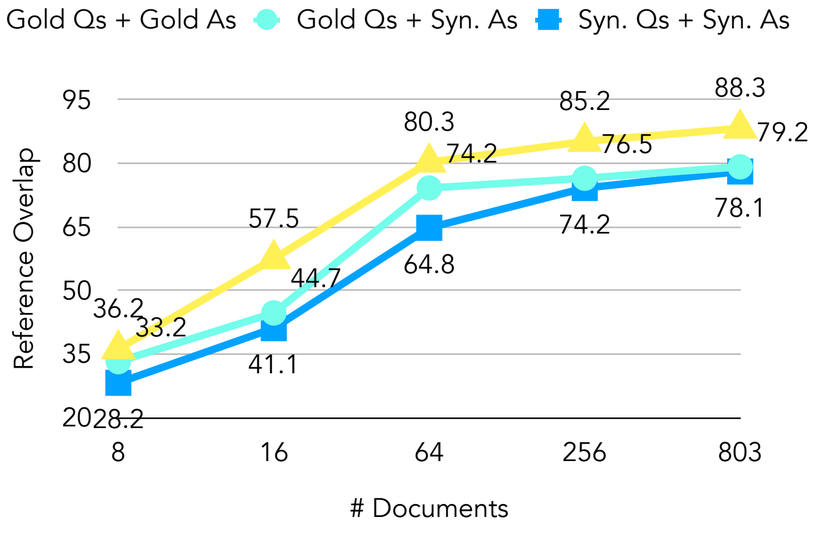


Question Distribution



Our method **significantly increases the proportion of challenging questions**, such as "Non-Overlap/ Answerable" questions, **from 4.2% to 24.2%.**

Synthetic V.S. Gold



Synthetic questions can be helpful, but **gold answers are important for optimum performance**!

Acknowledgement: This research is supported in part by funding from Optum. The authors would like to thank Optum for their generous support.