**Change after review: Paper\_05**

**Title: Context-Guided Medical Visual Question Answering**

* **Title making** 
  + **Author names, institute and emails added**
* **Acknowledgment and Disclosure**
* **Text update:** 
  + **Introduction (typos and errors)**
    - Moreover, we use a vision encoder of a pre-trained vision-language model as an image encoder
    - Reduce the number of *trainable parameters*
  + **Methods** 
    - Explain how h\_i is maintained: Where H\_i is the history information for the current question q\_i. For each image, H\_i begins as an empty list. After the first question is answered, the list is updated to include the question and its answer tokens. This process repeats after each subsequent question until the set has been completed.
  + **Results**
    - Clarify the version of hi-VQA in use: We want to clarify that when discussing results, "hi-VQA" refers to the updated results provided by the authors along with the dataset after the paper was published.
    - Metrics and result reporting: For a fair comparison, we use the same evaluation approach and script as hi-VQA and report our results accordingly using the conventional metrics Accuracy, F1, Precision, and Recall. The main results in table 1 correspond to the best run scores among multiple runs.
    - Clarify partially vs. Fully trained versions: the fully trained Context-VQA model refers to our model with all components, including the encoders, trained. With 232 million trainable parameters, this model takes an average of 1 hour and 48 minutes to train for a single epoch. However, due to the risk of overfitting on the relatively small dataset, we have strategically decided against pursuing this approach as it may not yield the most generalizable results.
* **Tables and figures**
  + Fig 1: Fix the medical errors and add a legend for the GPT API.
  + Fig 2: clarify “image-related” measurement
  + Fig 3: caption updated
  + Table 1: Fix F1 mistakes
  + Table 2: Mark in bold the best scores
  + Table 3: Fix mistakes: (232M instead of 230)
  + Table 4: include hi-VQA results