1. Research questions:
    a. First one is broad enough but also technical
    b. Second question - two questions in one
        i. Link second question to first one
        ii. Goal is to integrate language models (human input) with tabular data; don't need to address limitations with language data but explore them further - what can be integrated with LLMs and GANs
        iii. Changing data records is a limitation
        iv. Small dataset-performance of GANS will be low, evaluating = potential challenges
    c. Third question is also an evaluation question about synthetic data
        i. Read papers about this
        ii. Compare similarities between original and synthetic healthcare data, correlations or performance
        iii. Why do you want to generate synthetic data
        iv. Focus more on generation vs evaluation
    d. Fourth question valid RQ
        i. Linked to first one
    e. Throughout project will probably focus on 1-2
    f. First RQ should be limitations (identify problem, research gaps)
        i. 2,1,3,4 - two major and two minor
    g. Revise RQs after meeting
2. Dataset discussion:
    a. Can start from Kaggle dataset but in the long run requesting access to MIMIC dataset will be more beneficial - mirrors real world settings
    b. Only need 1-2 people to get access to it
    c. Test is very user friendly for MIMIC database
3. Schema:
    a. LLMs:
        i. Can download LLAMA - model is open source
        ii. LAION could also be interesting to look at
        iii. Go for free ones to start quicker
        iv. In report justify why you chose a specific LLM
    b. If doing time series can look at diffusion models - open source
    c. First challenge may be running the models
    d. To change feature names/values - ie: want 2,000 males:
        i. Generator focused
        ii. Sample focused - sample 20,000 and choose 2,000
        iii. Look at conditional GANS and start at sampling
        iv. Model by itself doesn't know logic, LLM will play a big role here
4. Limits of synthetic data is what the training/real world data is - ie: if BMI is up to 25 the synthetic data probably won’t go beyond 40
    a. Part of evaluation
5. Phase 1 presentation:

a. Literature review
b. Clear work plan - dataset, technology, lit review, risk analysis, how to connect everything
c. Be concise for the report
    i. Can talk about what is synthetic health records data