Things to Consider for your Data Collection Plan

As you make a plan to collect data for a custom keyword spotting application, there are several important data engineering questions to consider in designing your data collection and testing scheme:

- Who are the anticipated end users?
    - What will the age range be?
    - What languages will the user be speaking? Will the users have accents?
- What are the goals for using the application? How will this impact the requirements / needs of the ML model's performance?
    - For example, will the user be trying to turn specific lights on/off vs. turn the thermostat up/down vs. arm/disarm a security system vs play your favorite playlist from a smart speaker, etc.?
- What false positive or false negative rates can the application handle / what would the consequences be of such an event?
- In what environments will the users employ the application?
    - How much background noise do we expect?
    - How far will the users be from the device, and from noise sources in the environment?
    - Will these values vary over time?
- In what situations will the users use the application? Will the user's speech sentiments vary?
    - Will the user be stressed vs. calm vs. panicked?
    - Will the user use different volumes of voice (whispered/normal/loud/shouted)?
    - Will different users have different sentiments?
- How do all of these previously mentioned factors affect how much and what kinds of target, non-target, and imposter/adversarial data you will need to train with?
    - How likely is it that these wake words will be triggered unintentionally during conversation?
    - Who/how many people will be talking in the application environment?
- Given all of these previously mentioned factors what will you collect?
    - What custom wake words will you select?
    - Are the background noise samples from the Speech Commands dataset sufficient for model robustness? Would it be helpful to collect additional background noise samples from specific environments?
    - Do you need to collect other words to ensure the model learns the difference between them and your wake words?

    - How much data will you need to collect? We all know more is better (usually), but you also live in the real world with time constraints so how much do you think will be enough?
    - In what environments will you collect these samples?
- How will you test your model?
    - What words will you try? How many times will you try these words? Will any of them be adversarial? Remember even current production systems struggle with close words at times!
    - Who will test the model? How diverse will the testers be?
    - Will the testers vary their speech sentiments?
    - Where will the testers be in relation to the device?
    - In what environments will the testing happen?
- Finally, based on your initial testing, how will you improve your results:
    - What didn't work the first time?
    - Will you need more or different data?
    - How can you iteratively improve your results?