|  |  |  |
| --- | --- | --- |
| Name & Dept: Use Case:  Date: | | |
| 1. Data: *Do you have data to support this use case? If so, describe any data that you can use as* ***target*** *data, i.e., data that an AI model can process for this use case. Do you have data that can* ***train*** *an AI model (data that already has the features you'd like the AI models to be trained on) and data that can be used to* ***verify*** *that the output of models are correct?* | 5. Pre-processing: *Describe the steps and any transformations used to create the dataset. E.g., text from a digitized document that was OCR’d. When and with what tools was the data transformed? Was the data cleaned or normalized? If so, how?* | 8. Structure & Storage: *How is the data structured? E.g., in XML, CSV, unstructured text, etc. Does the structure follow any standards? If so, what are they? How and where is the data stored?* |
| 1. Composition: *What is represented in the data? Describe the language, time period, genre and other descriptive information about the intellectual content of the data.* | 9. Characteristics, Patterns, Labels: *What are the characteristics or patterns the AI system will detect in the data? Describe the data elements the AI will predict or output for this use case?*  *How were the patterns labeled? Are they naturally occurring, did experts label the patterns, or did unskilled or crowdsourced staff or volunteers label the data elements? What was the incentive structure for the labelers, if any.* |
| 1. Compilation: *How was the dataset compiled? E.g., via API query or bulk download? When? With what tools or expertise?* | 1. 6. Data provenance: *Describe the relevant background on where the data comes from, why it was created, by whom, where, and when. Include any version information and if the data is used in other systems* |
| 4. People: *Who is depicted in the data? Is there any PII in the data? Are people depicted in the data described in a potentially outdated or harmful way? Are the people depicted in the data aware their data will be part of an AI system?* | 7. Restrictions/Controls: *Who owns the data? What is the Copyright status of the data? Is the data restricted by privacy, confidentiality, license, or other terms?* | 10. Sampling and known biases or imbalances: *What sampling method was employed to create the dataset, if any? What are the known biases in the dataset? E.g. language, geographic, demographic, historic.* |