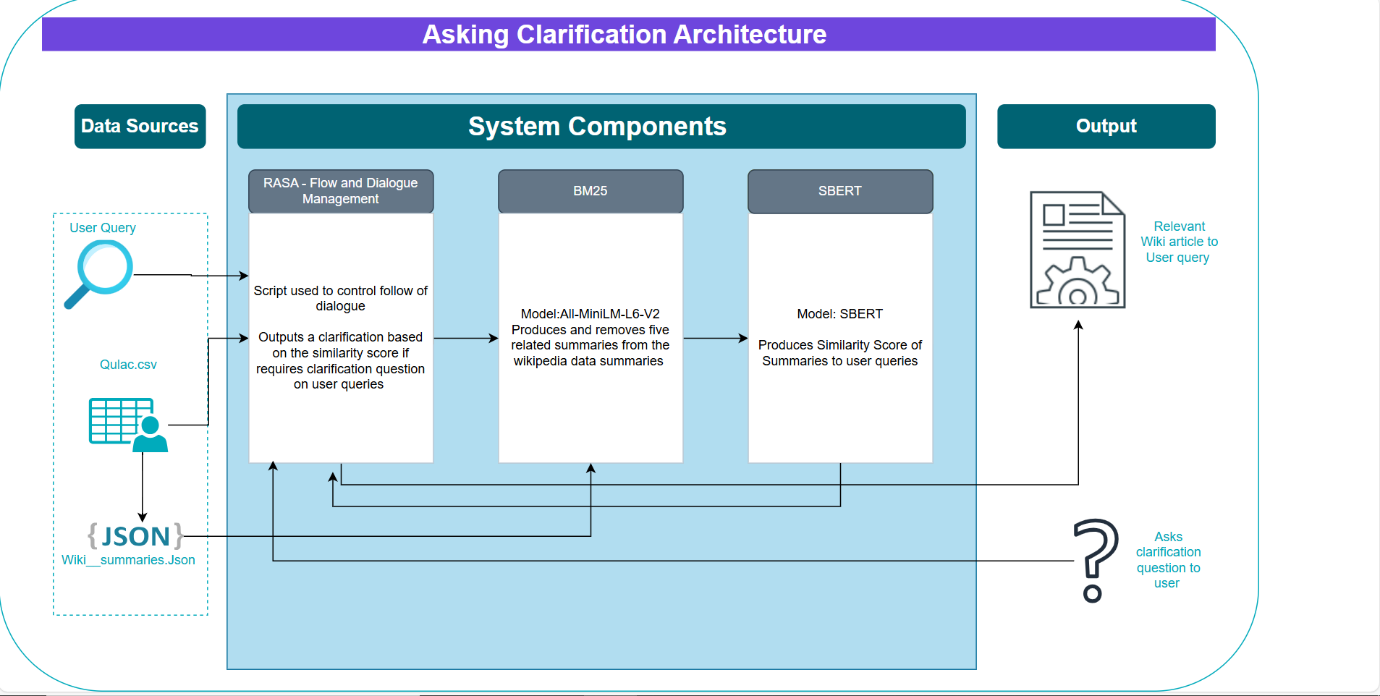
**Methodology**

The system for the clarification model was constructed was through a series of key component parts: User query, RASA for flow control and dialogue management, BM25 for pruning the queries on the Wikipedia Dataset, BERT for constructing a Sentence Similarity Scores. The dependencies and relationship of these components can be shown below.

**System Architecture   
**

**Qulac Dataset**

The Qulac data was one of primary datasets for this system. One of the first elements was to shape the data appropriately to work for the other elements within our systems. The transformations for cleaning this dataset were performed in Python using the Pandas library and the resulting CSV was stored within a Google Drive directory.

*Clean Qulac Schema*

|  |  |  |
| --- | --- | --- |
| Columnar Name | Type | Description |
| Index | String | Facet Index based on word string |
| Category Name | String | Name of the Broader Category of the Dataset |
| Facet Description | String | Facet Description |
| Col 1 | String | Response 1 |
| Col 2 | String | Response 2 |
| Col 3 | String | Response 3 |
| Col 4 | String | Response 4 |
| New Index | Int | Index counter for every row in the table. |

One of the most important features of our Qulac Dataset was the facet descriptions as it was used to derive the Wikipedia Summaries.

**Wikipedia Summaries**

The Wikipedia summaries were generated using a series of GET responses contained within a python script. The aim of this method was to generate



**RASA**

**BM25**

**SBERT**