Ultimately, this project successfully demonstrated a functional conversational search system that takes user queries as input, and outputs clarifying questions or satisfactory wiki summaries. By combining a retriever model (BM25) with a semantic reranker (all-MiniLM-L6-v2) to form a comprehensive summary retrieval pipeline, and a fine-tuned SBERT model for clarification questions, the system was able to deliver relevant responses or appropriate follow-up questions to the user. These interactions were handled through the Rasa dialogue manager, which acted as a middleman ensuring smooth conversational flow and intent management.

Evaluation results showed that the retrieval pipeline achieved strong performance on matching queries (92% accuracy, 95.56% F1 score, and 0.9594 average semantic similarity). This confirmed its effectiveness in returning relevant summaries when sufficient information was provided. On ambiguous queries, the system produced confident predictions only 42% of the time with a lower average similarity of 0.3584, thus demonstrating excellent uncertainty handling. For non-matching queries, the system was correctly cautious in 80% of cases.

The SBERT clarification handling also showed promising results, ranking the correct question with confidence 84% of the time, with an average similarity of 0.563 across 75 queries of differing nature. It also showed its ability to give uncertainty (12 queries below the confidence threshold) which, supports expected behaviour in ambiguous contexts.

A key component enabling the system’s interactivity was the Rasa dialogue manager. Rasa handled user input, tracked conversation history, and coordinated the conversational flow between the user and the model seamlessly. Rasa helped the system mimic real conversational behaviour, therefore making it an apt tool for managing interactive search.

Although limiting factors such as the size of the dataset sourced from the Qulac dataset, and the fixed set of clarifying questions constrained the system’s general ability somewhat, the pipeline still proved effective within its scope. Additionally, the SBERT model was only fine-tuned on a small subset of the data, which may have lessened its ability to generalise across rare query types. Fundamentally however, the system achieved its primary goal: to decide when clarification is needed and deliver either a relevant summary or an appropriate clarifying question. Within a lightweight framework, it demonstrated reliable performance and the potential for real-world applicability in conversational search.