# **Challenge: Context-Aware QA Under Token Constraints**

## **Objective**

Build a question-answering system that selects and presents the most relevant context to a local language model under a strict context token limit. The LLM can only “see” a small slice of the knowledge base, so your system must make smart decisions about what to show.

## **Provided Materials**

You will receive:

* A fictional knowledge base (docs) consisting of 10 short .md files
  + Domain: sci-fi technical manual
* A file containing 5 natural-language questions (questions.json)
* A token budget (max 1024 tokens of context per question)

## **Your Tasks**

### **1. Chunking**

Split the documents into meaningful text chunks and design a way to retrieve and score them while staying within budget.

### **2. Answer Generation**

You may use a local small LLM (1.7B-2B).

**3. Evaluation & Justification**

For **each question**, choose what to show and how to justify why your method works well or could fail. You may elaborate on what may or may have been helpful to facilitate evaluation.

**Deliverable**

* main.py or notebook.ipynb - to be delivered at least 1 day before interview
* a powerpoint to explain your approach, results, model choices, difficulties, areas of improvement etc.