

**Pattern Recognition**

**Project Report:**

RAG with LLM dataset medical books in pdf format

**Group members:**

# **FA21-BCS-066 HUSSAIN ALI**

# **FA21-BCS-080 ALI SHAN**

**Invigilator:**

Dr.Samia Riaz

**Project Scope:**

**Objective:**

The objective of this project is to build a **Retriever-Augmented Generation (RAG)** system that can answer medical questions based on information extracted from medical books in PDF format. The system will integrate a retrieval-based model to find the most relevant pieces of information from a large corpus of medical data and a generative language model to synthesize responses based on these retrieved documents. This system will be designed to help medical students, professionals, and researchers quickly access accurate information from a vast range of medical books and references.

**Input:**

The input to the system will consist of medical books in PDF format, which may cover a variety of topics such as anatomy, pharmacology, diseases, treatment methods, and diagnostic procedures. These PDFs will first be processed into text, which will then be split into smaller chunks to facilitate easier retrieval and better performance when searching for relevant information. These chunks will be stored in a vector database, where they can be efficiently searched using embeddings to retrieve the most relevant information based on a user query. The system will return answers that are both accurate and detailed, relying on the generative capabilities of a fine-tuned language model.

**Goal:**

The goal is to create a robust system capable of answering a wide array of medical queries with high precision and reliability, using content directly extracted from the medical books.

**Dataset Link:**

<https://drive.google.com/drive/folders/1twxYrqAxzDcimg1iFKmsfByA3A9B-6IT?usp=drive_link>

**Dataset Structure:**

The **medical books** in the dataset are in **PDF format**, and they will first need to be processed to extract text. Once the text is extracted, it will be split into smaller **chunks** to facilitate easy retrieval. These chunks will be converted into **vector embeddings** using a model like **sentence-transformers** or **bioBERT** to ensure that semantic meaning is captured.

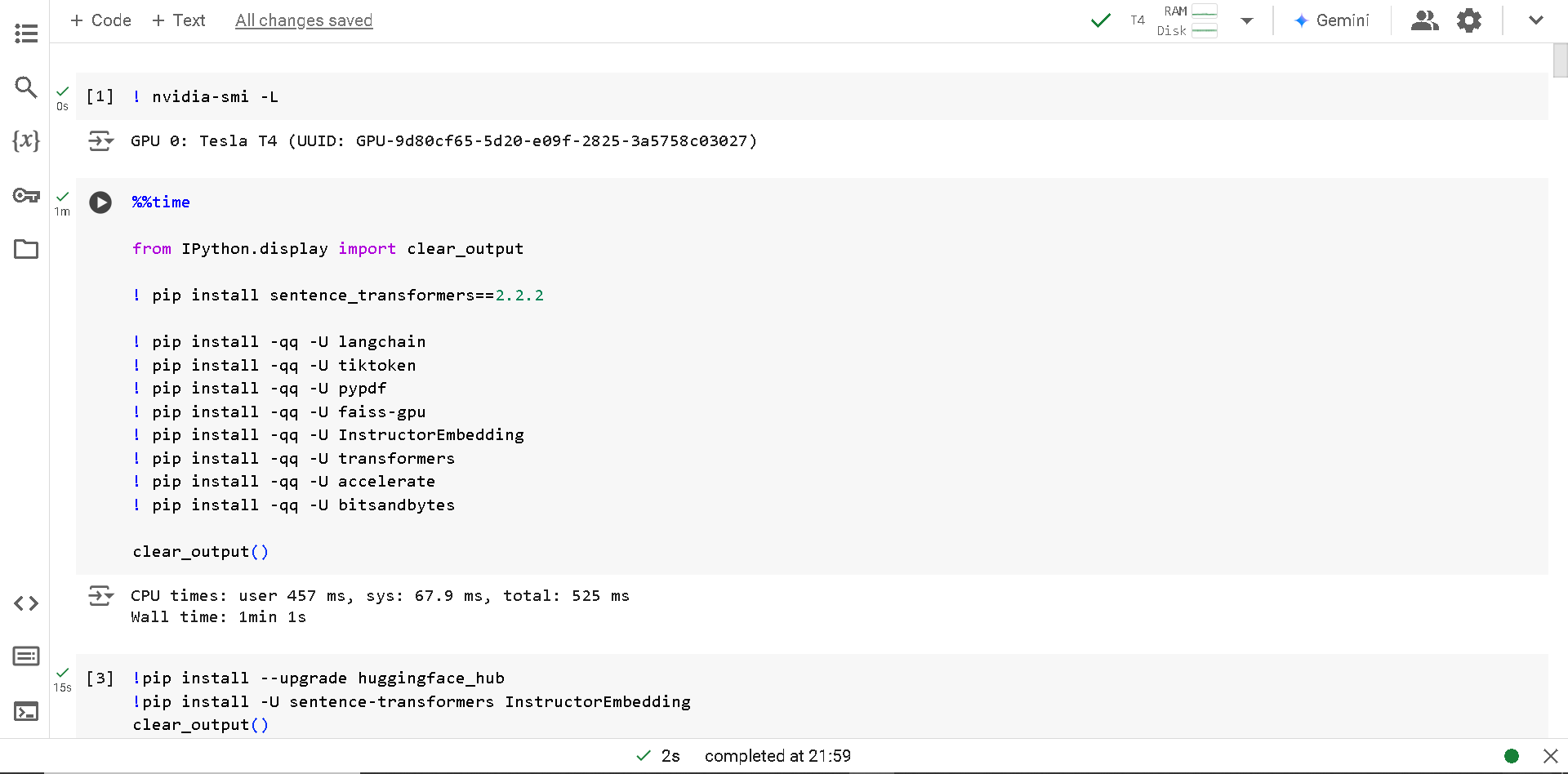
The **vector database** will store these embeddings, and when a user asks a medical question, the system will retrieve the most relevant chunks based on the query's **semantic similarity** and generate a comprehensive answer using the **generative model** (like **Llama-2**).

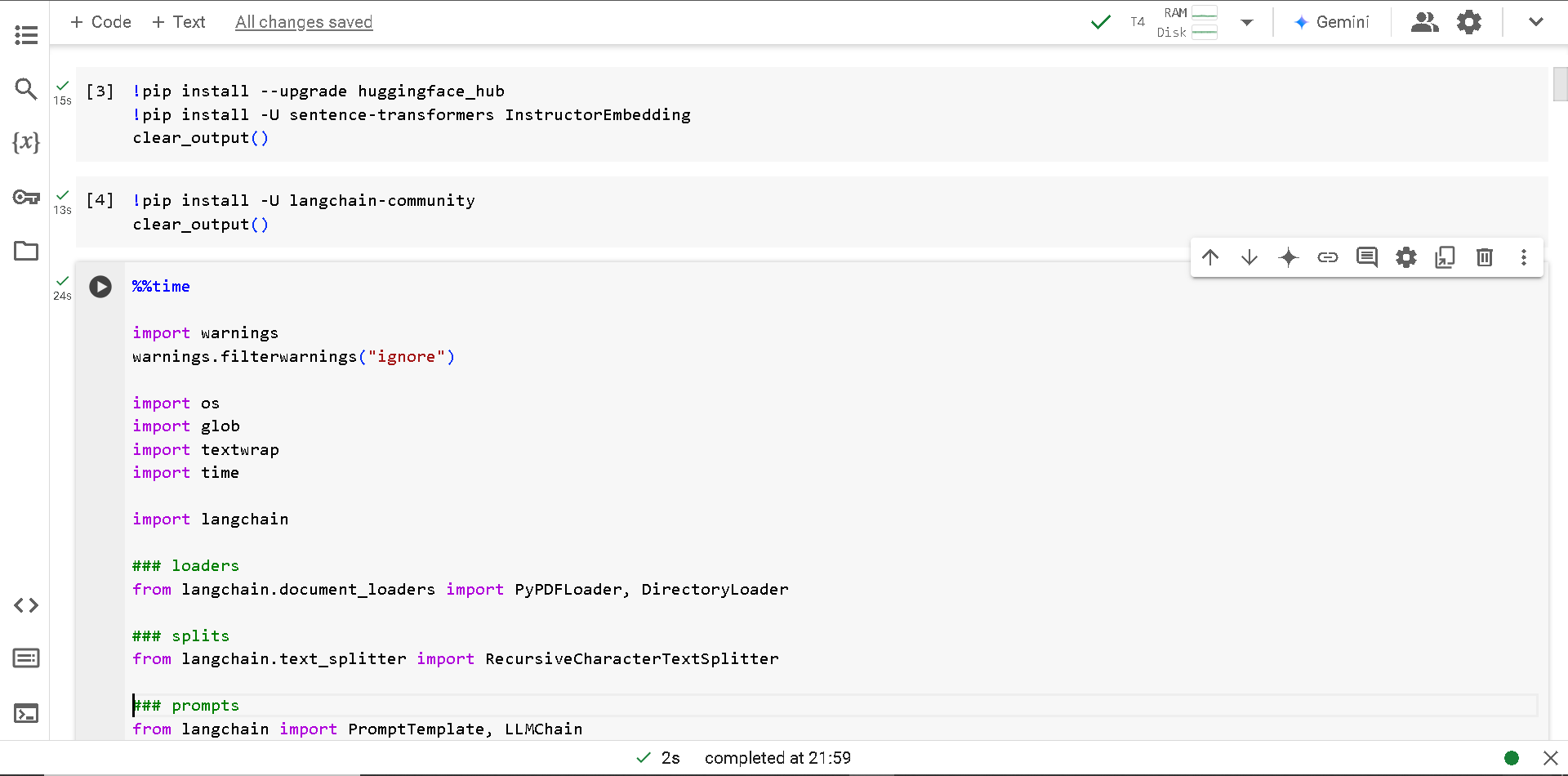
This dataset is essential for training the retriever and fine-tuning the generative language model so it can provide accurate, context-aware medical information.

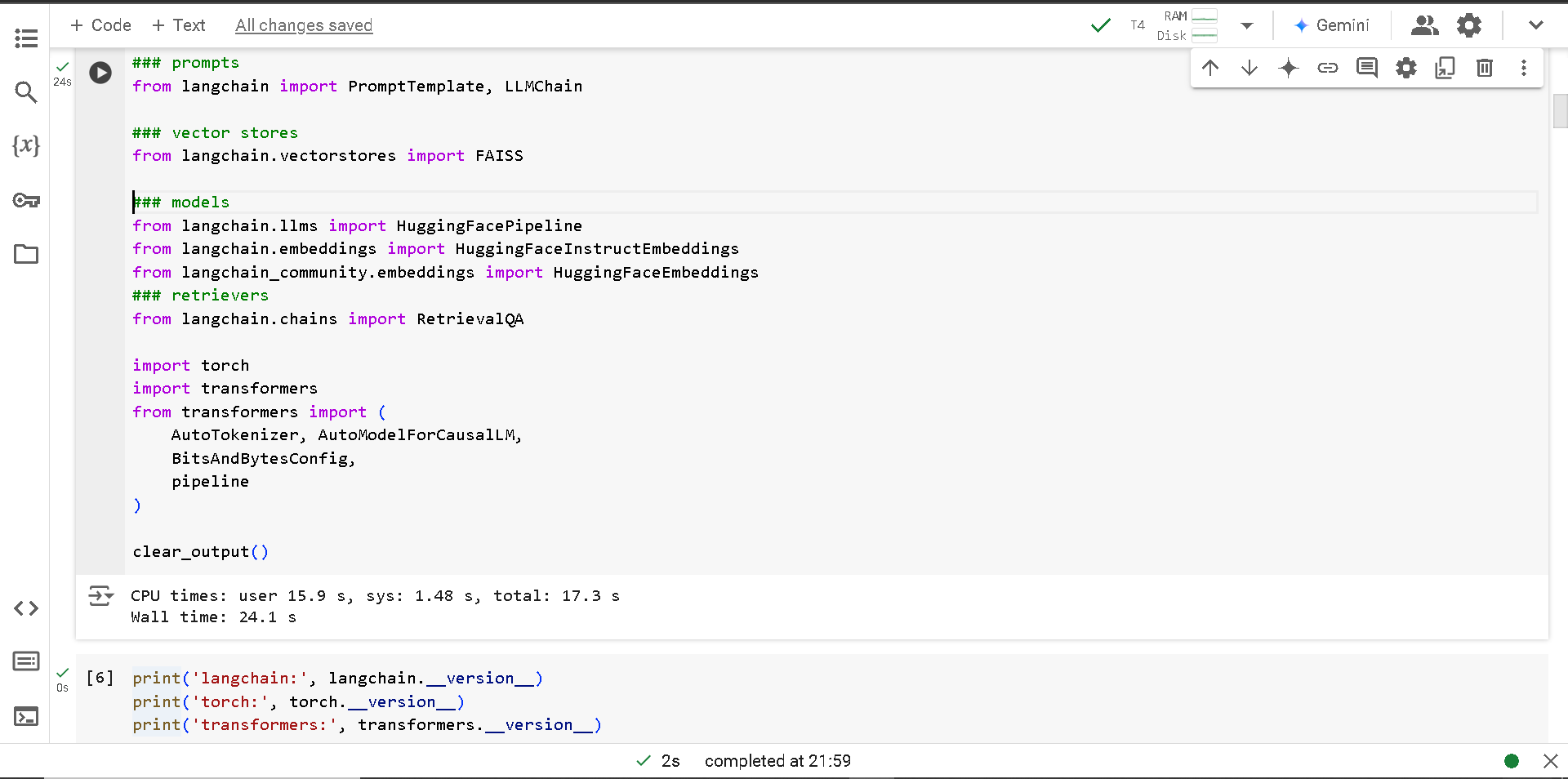
**Conclusion:**

This project illustrates the potential of **RAG systems** in the healthcare domain, providing a powerful tool for information retrieval and natural language generation. By using advanced machine learning techniques, it offers an efficient and scalable solution for medical question answering, which can significantly benefit medical students, healthcare professionals, and patients seeking reliable information.

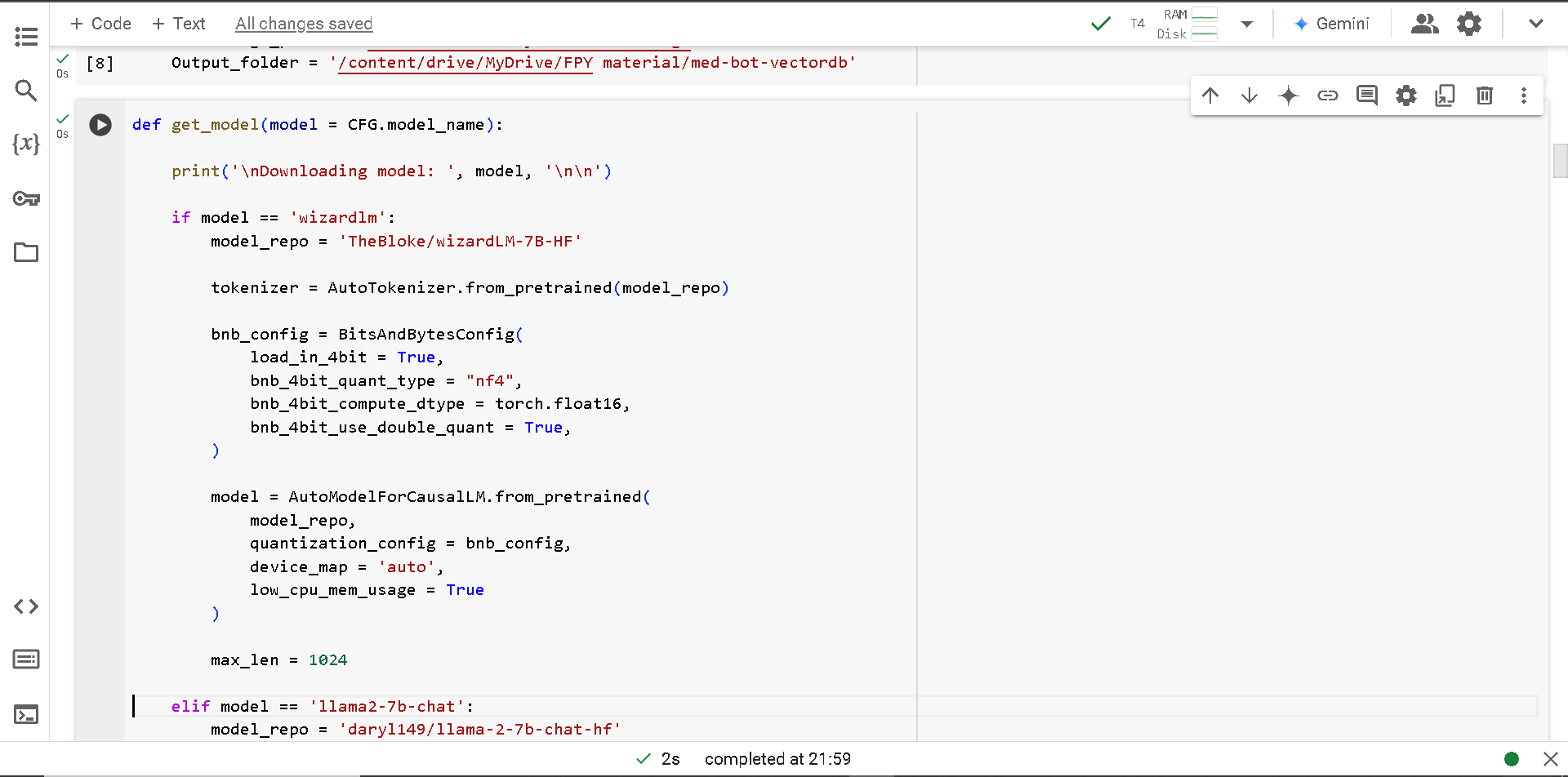
**Code Screenshots:**

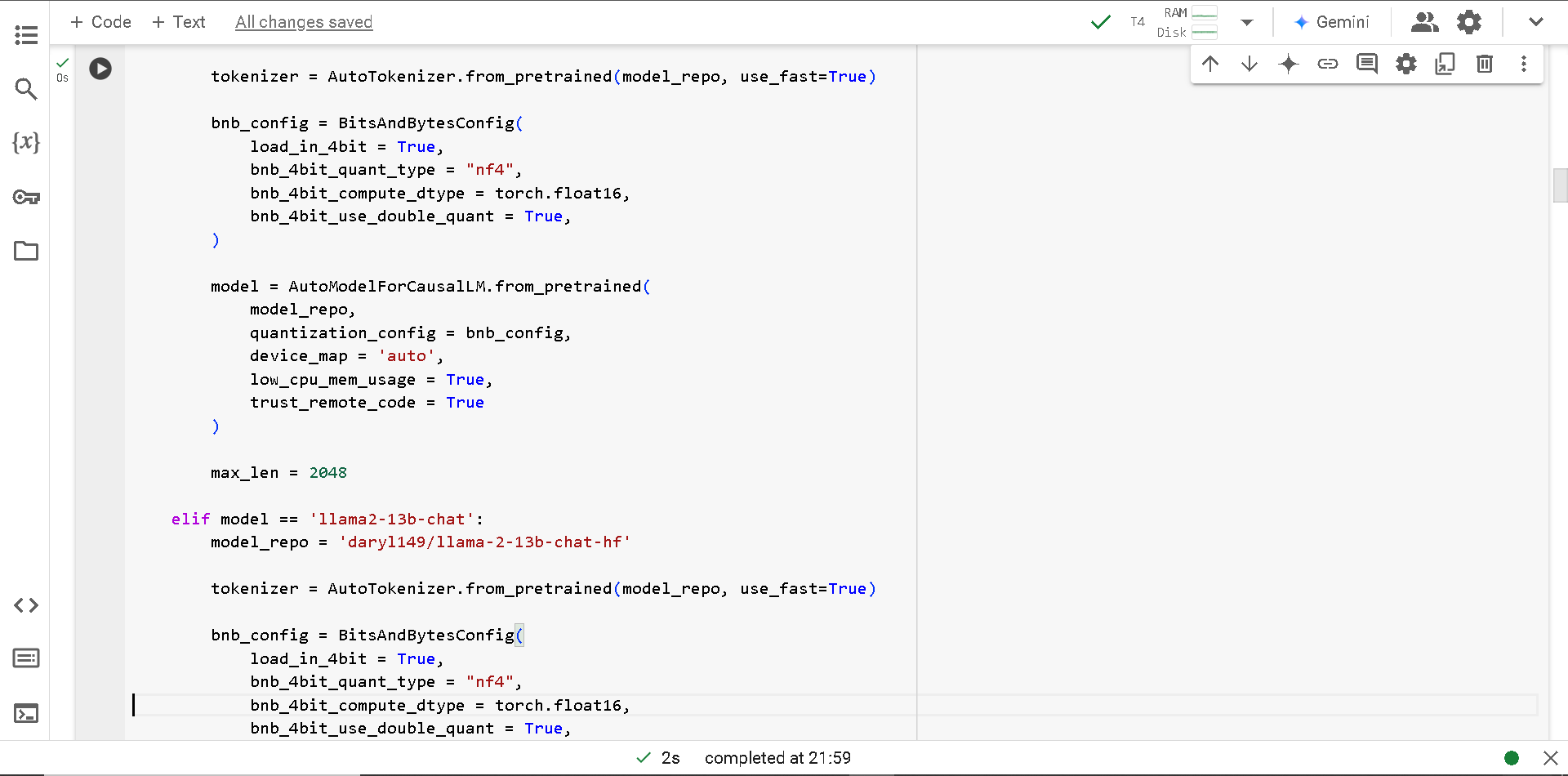


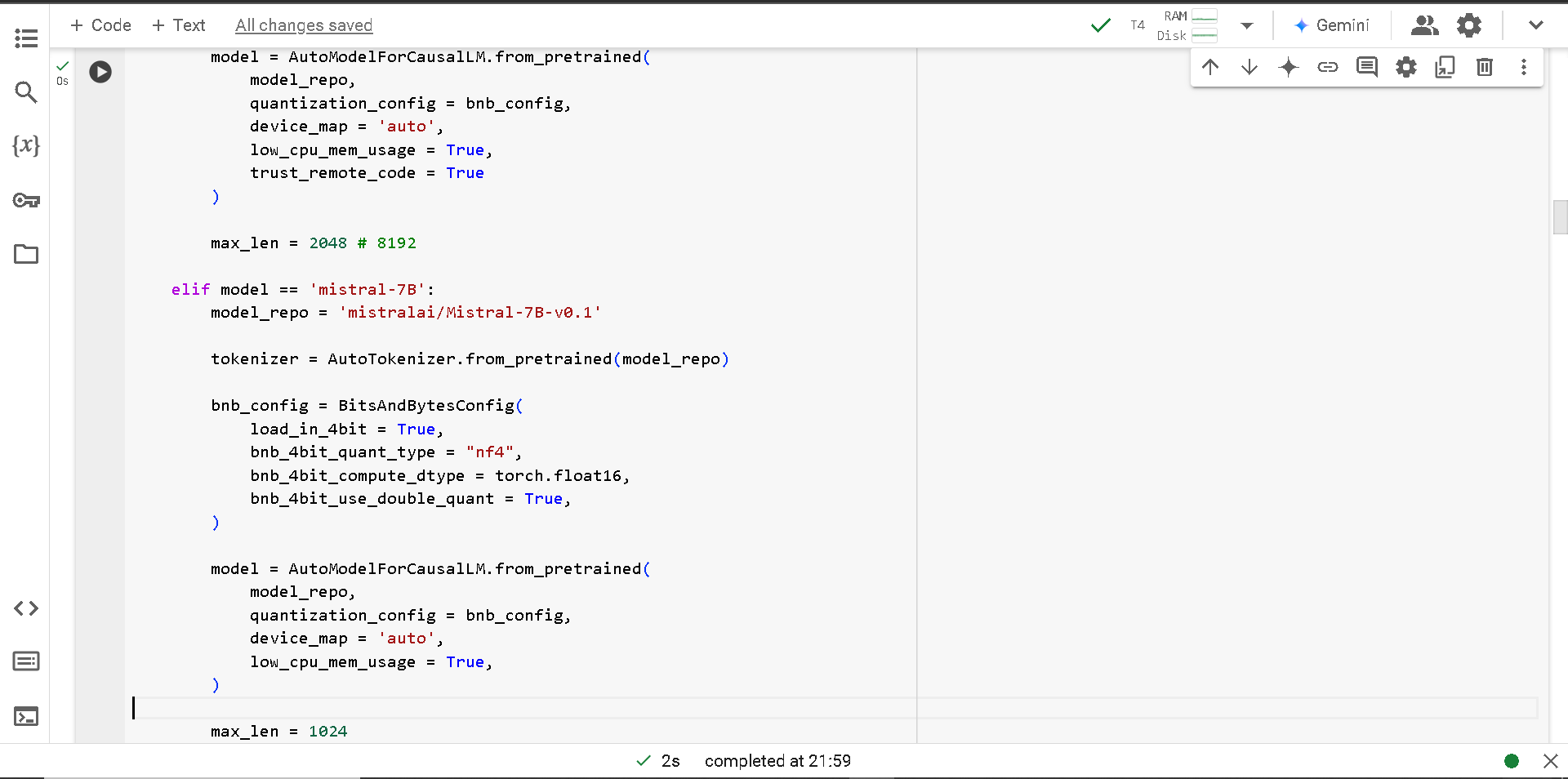


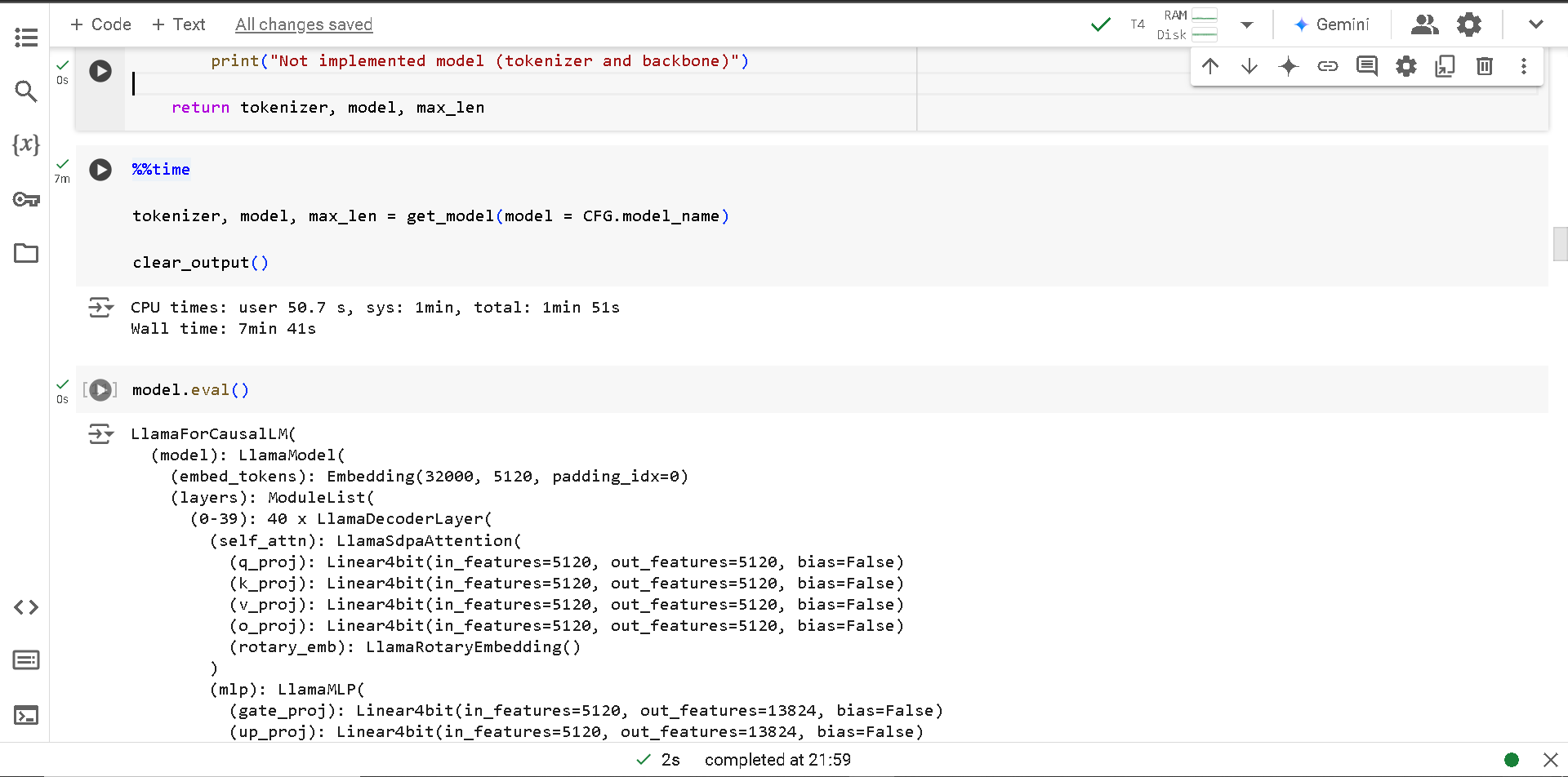


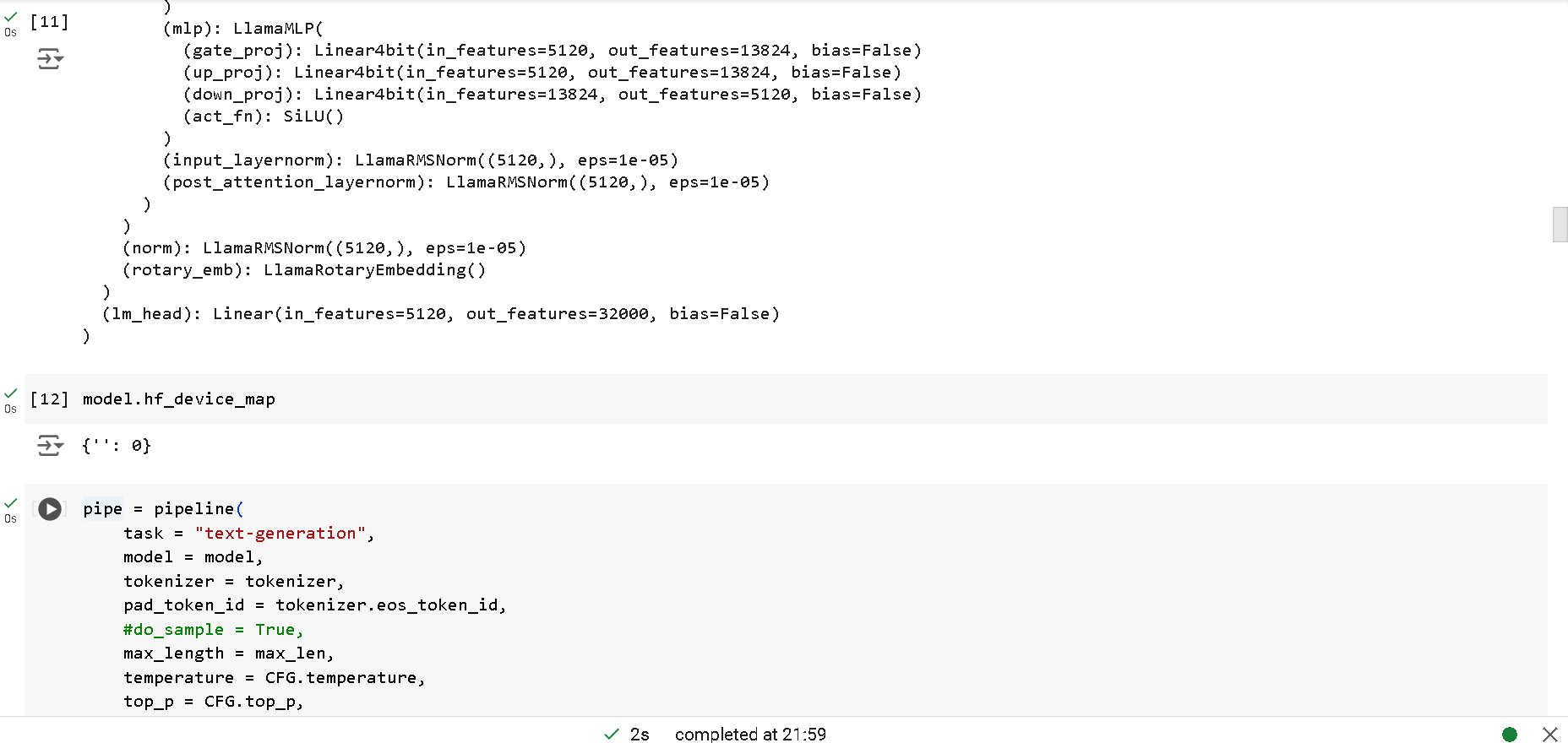


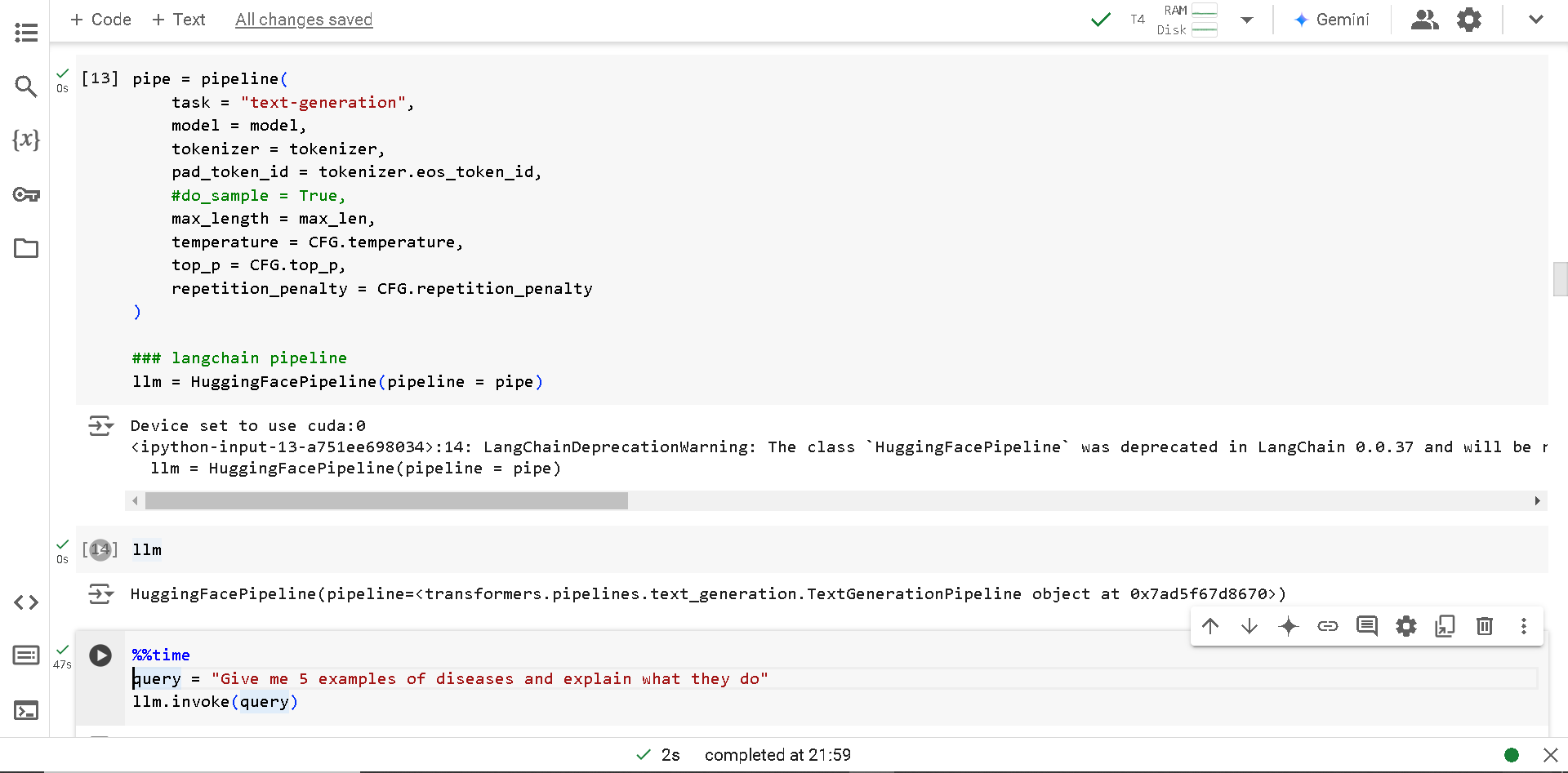


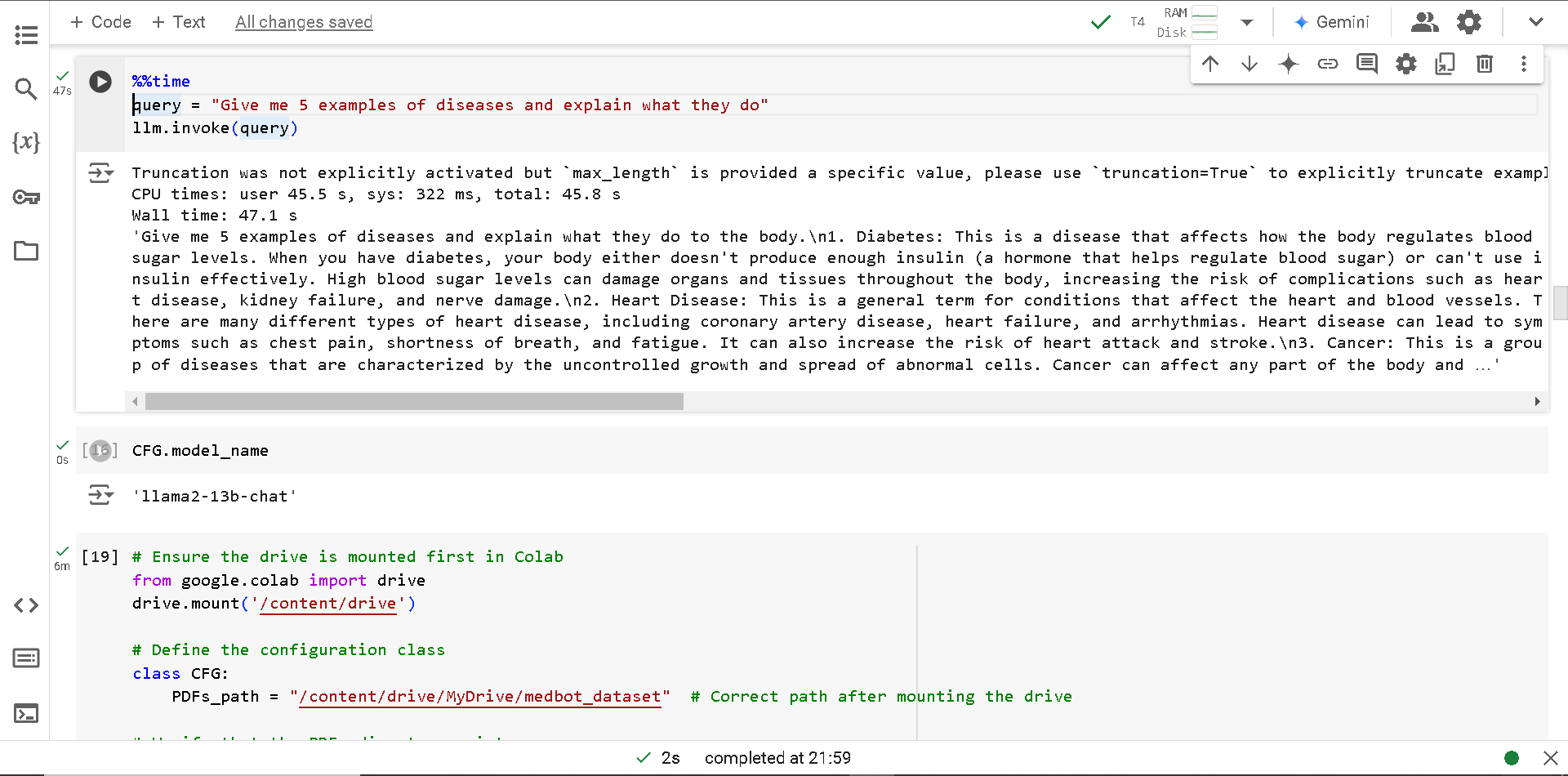


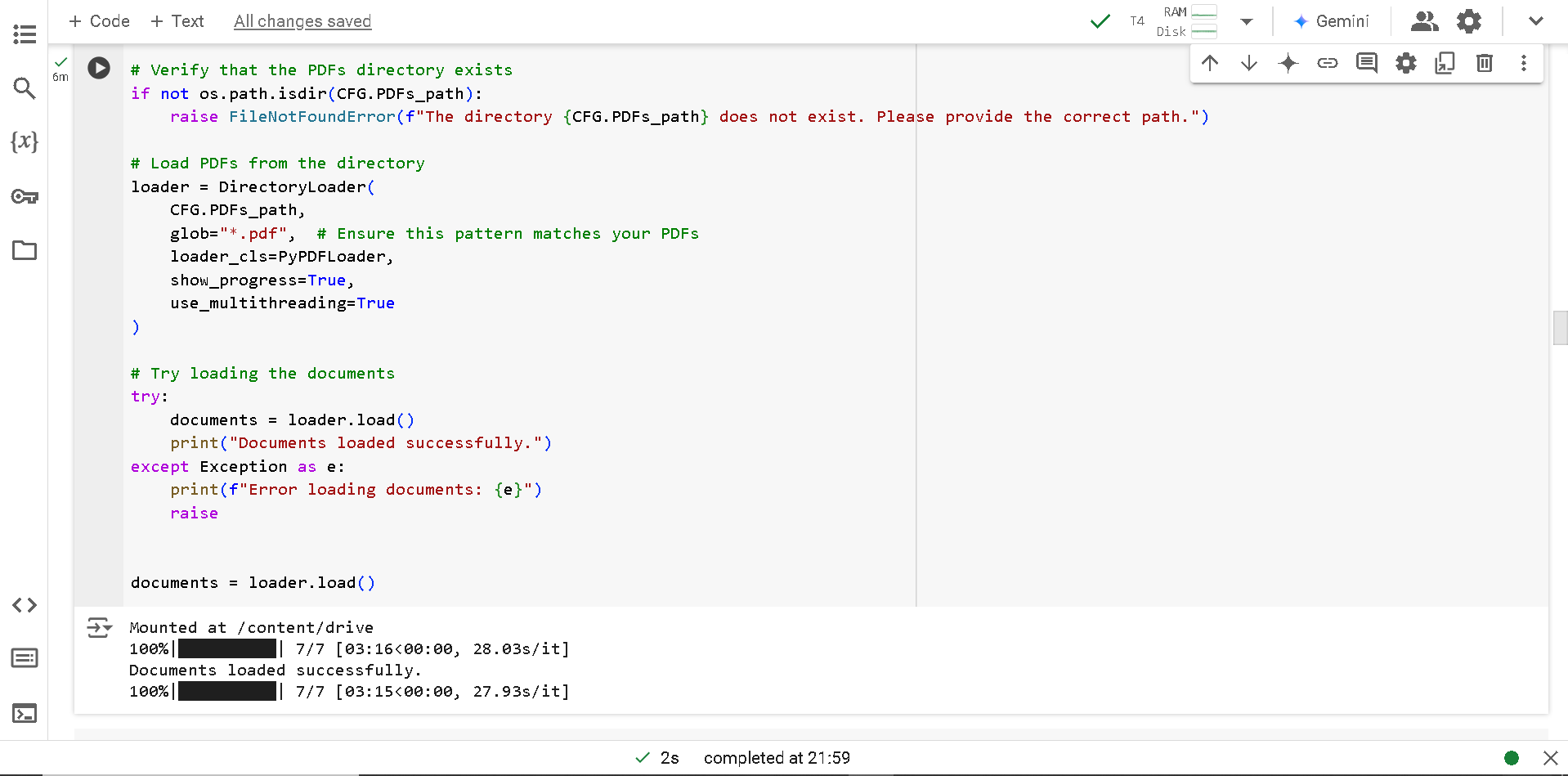


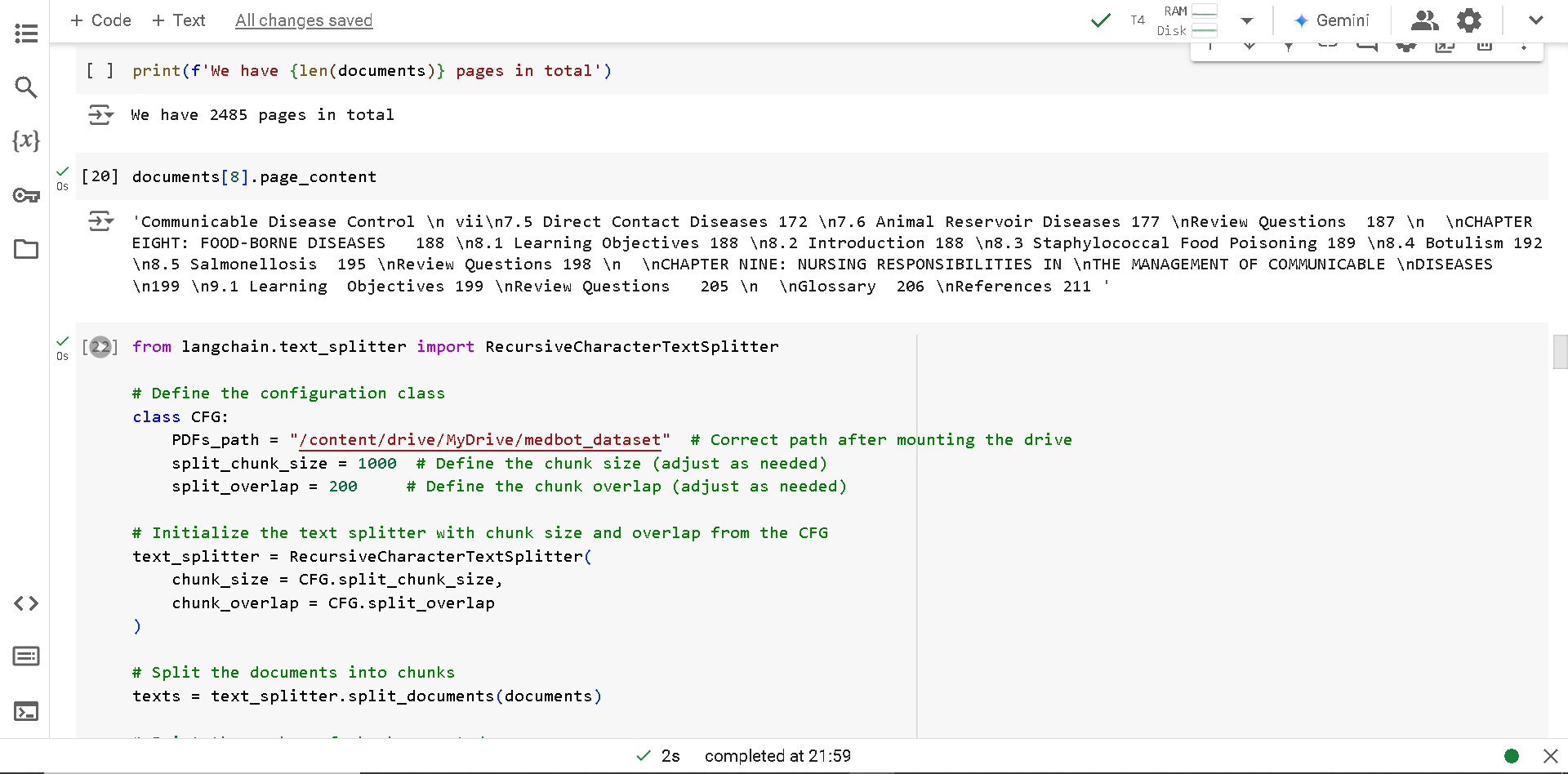


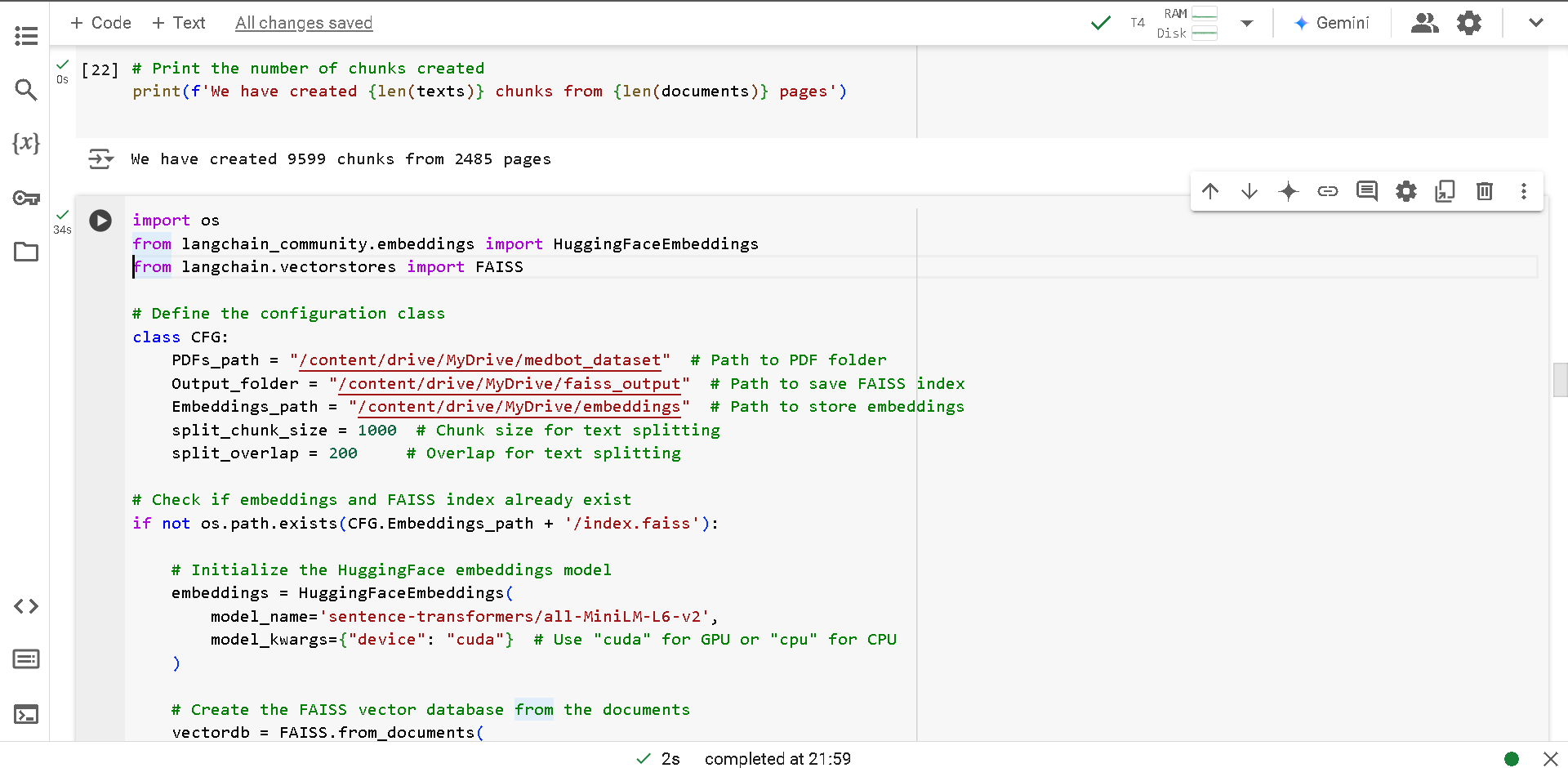


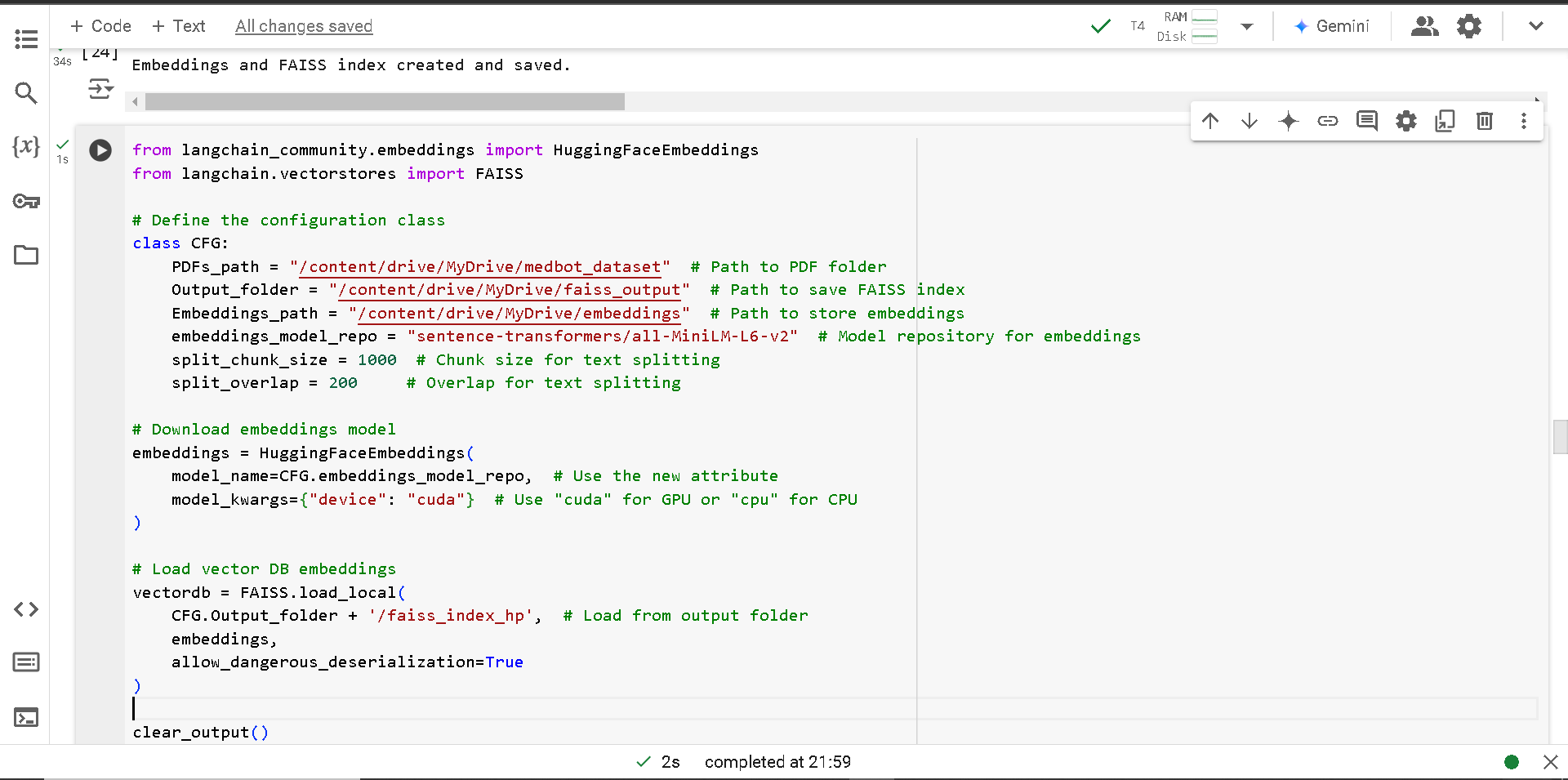


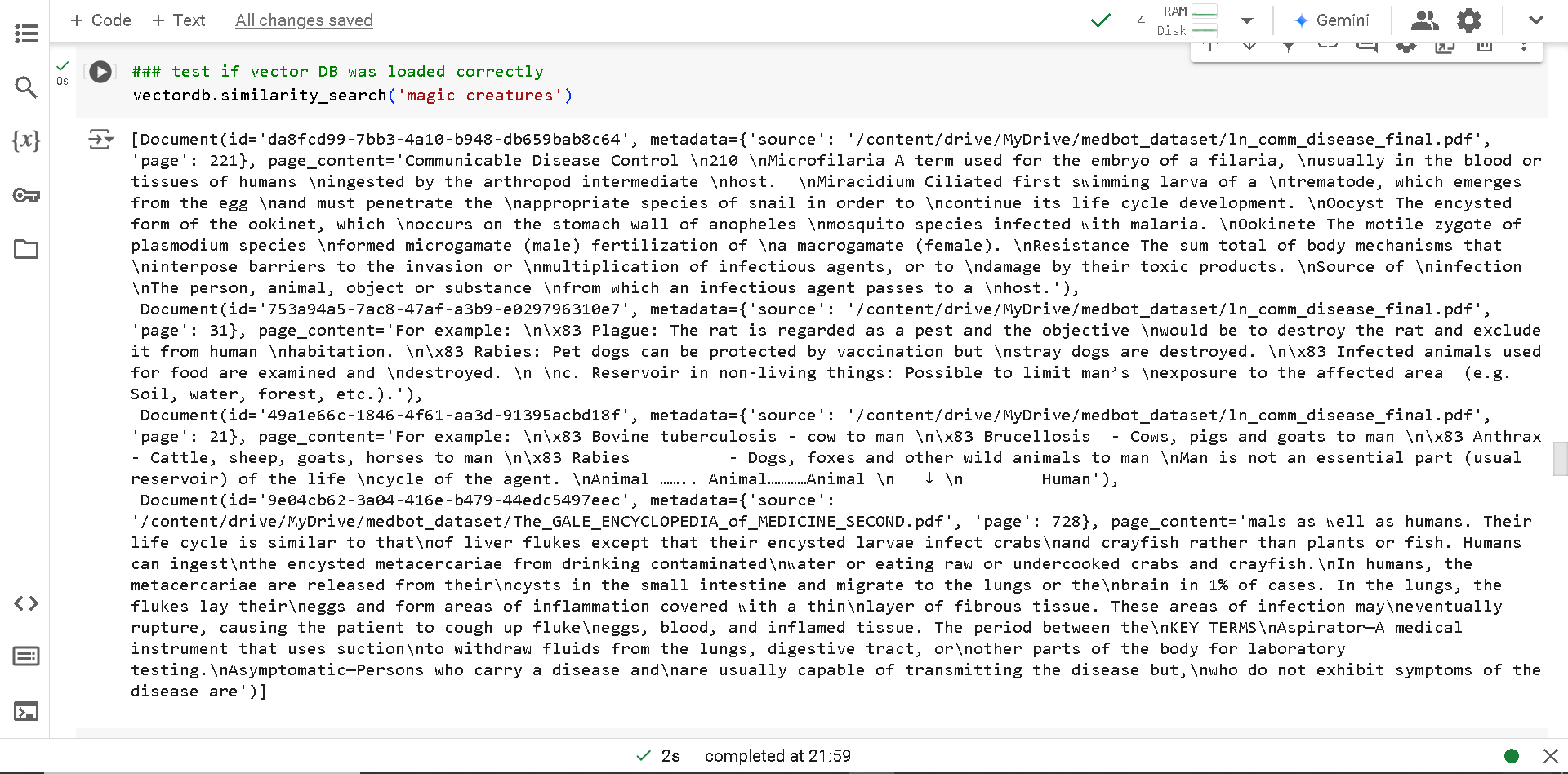


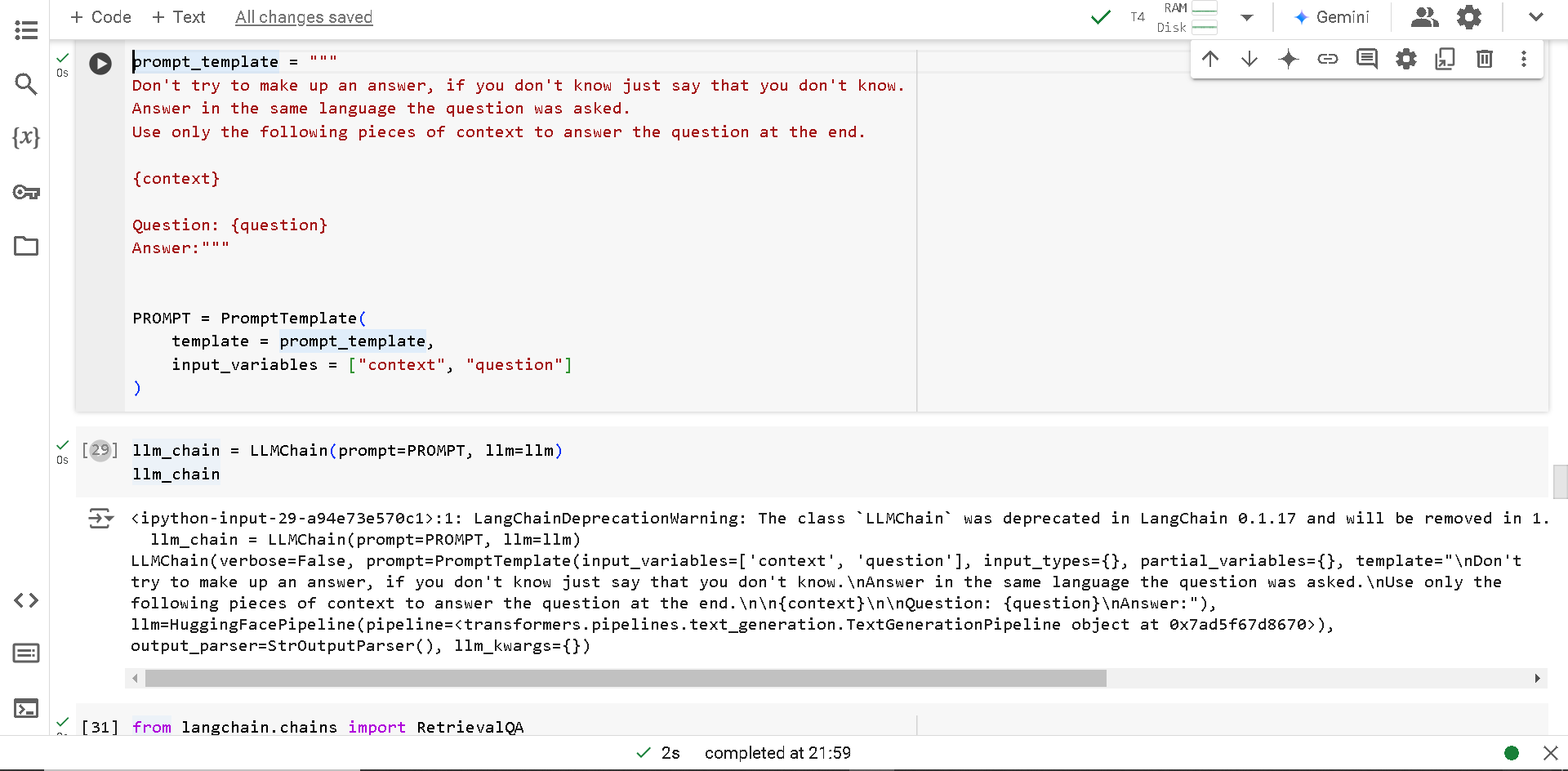


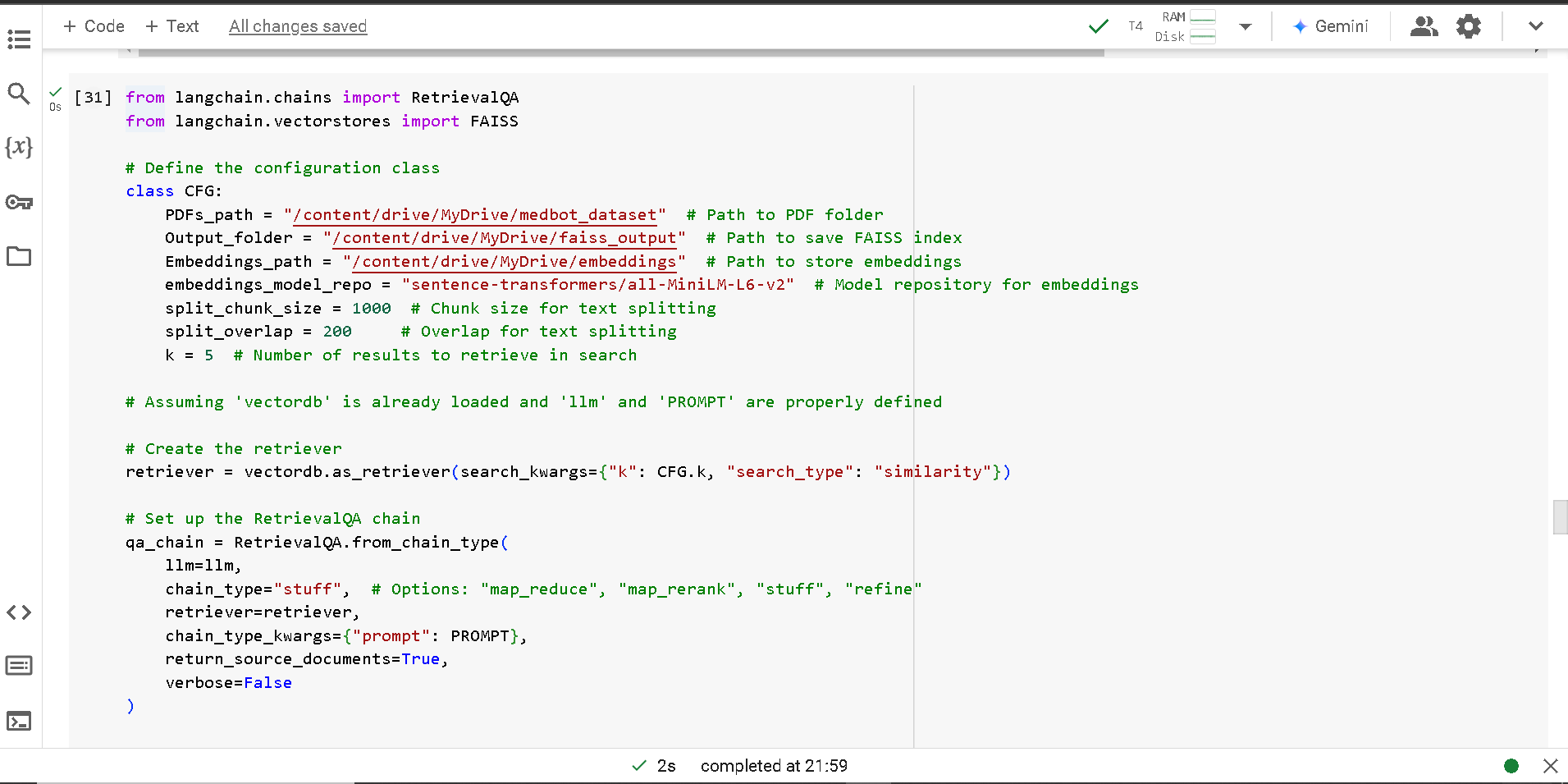


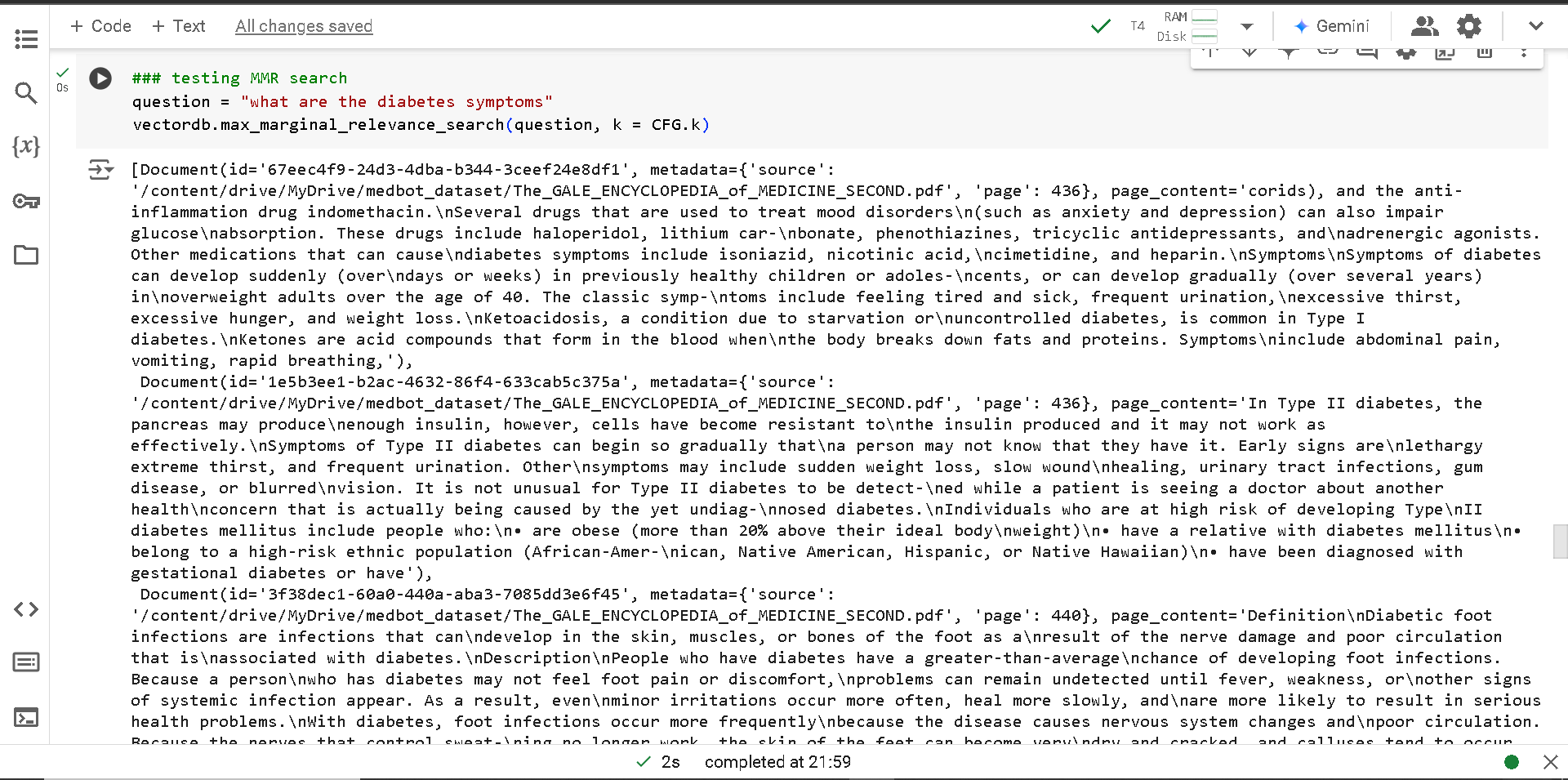


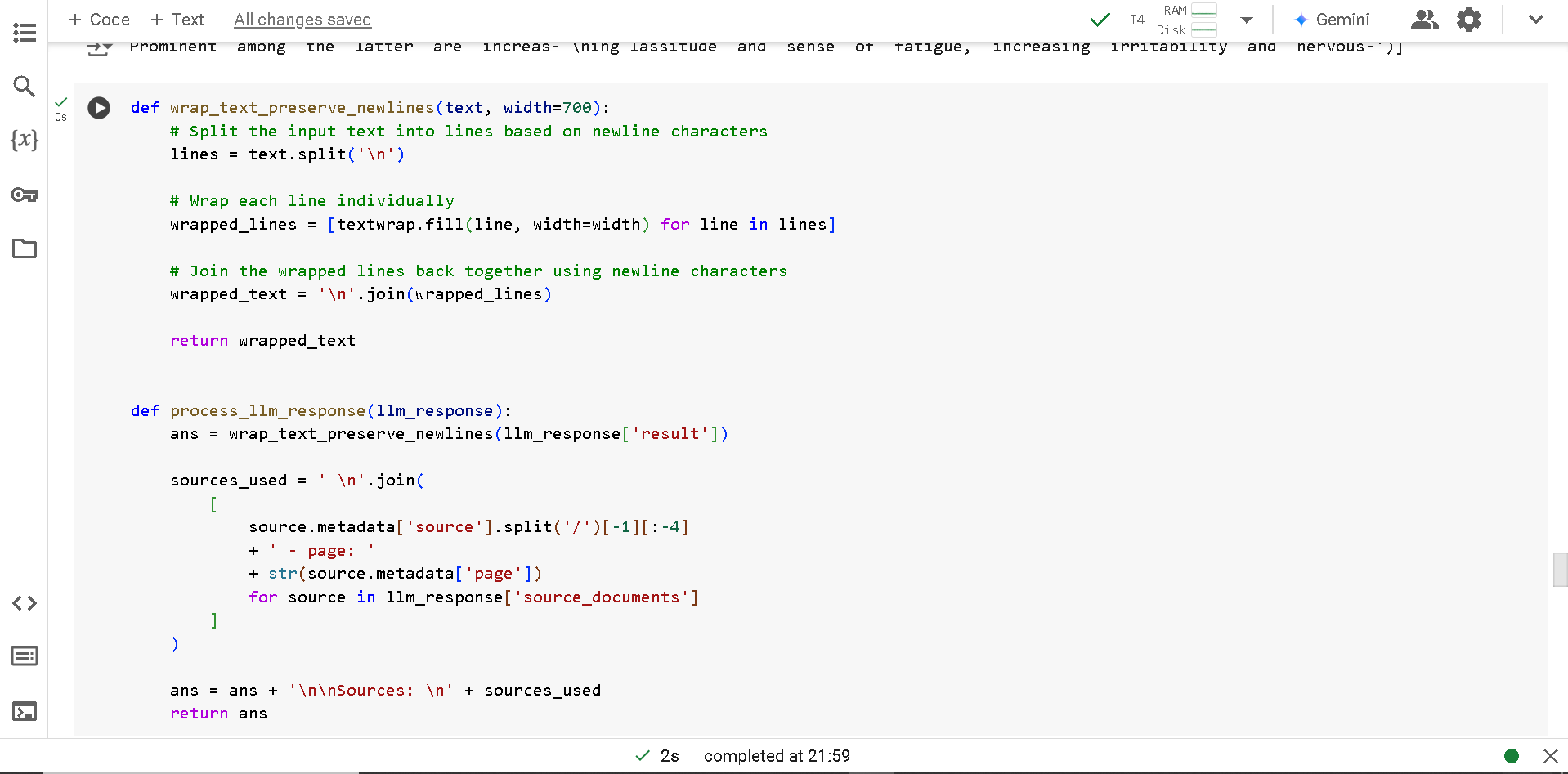


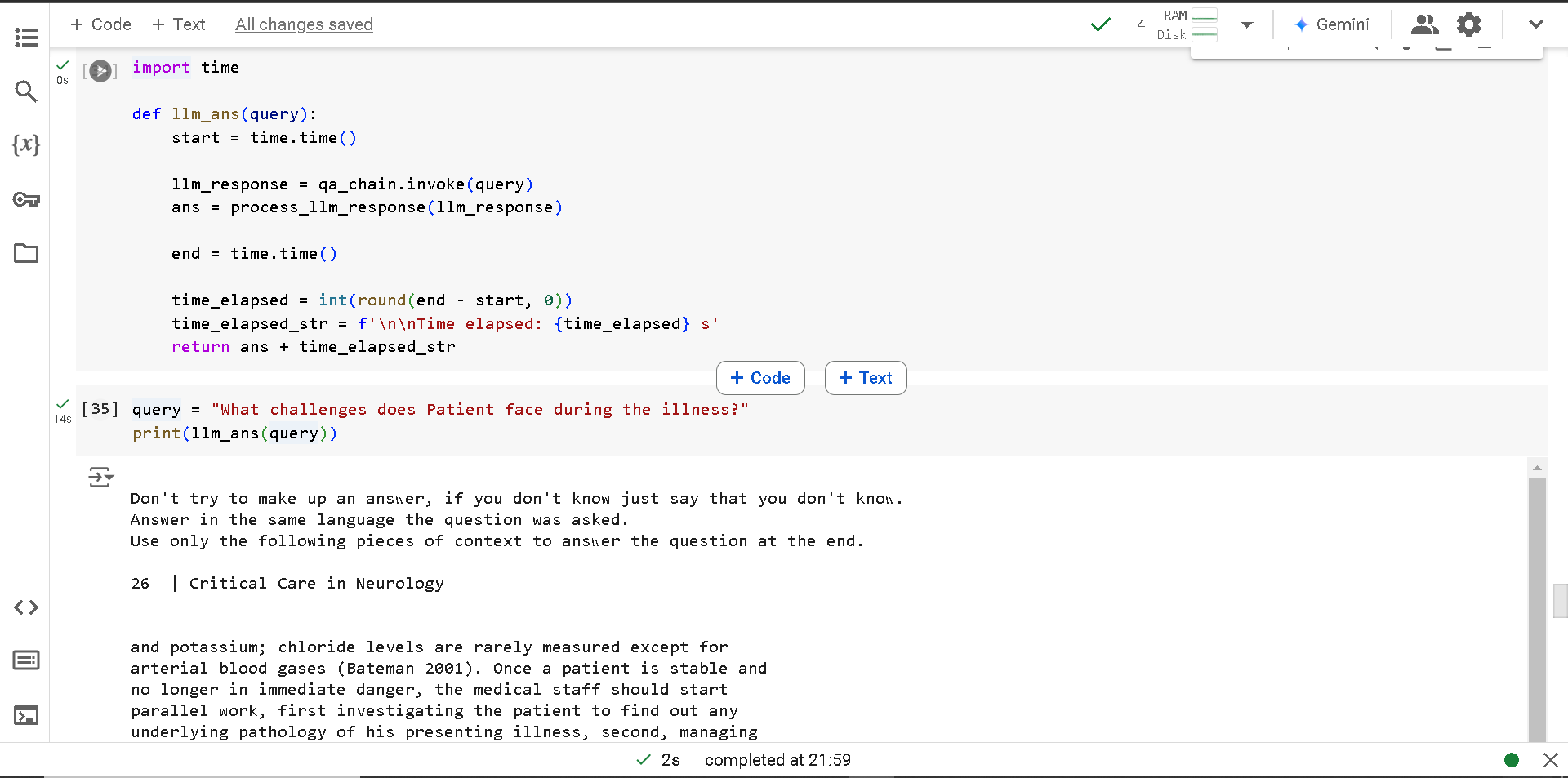


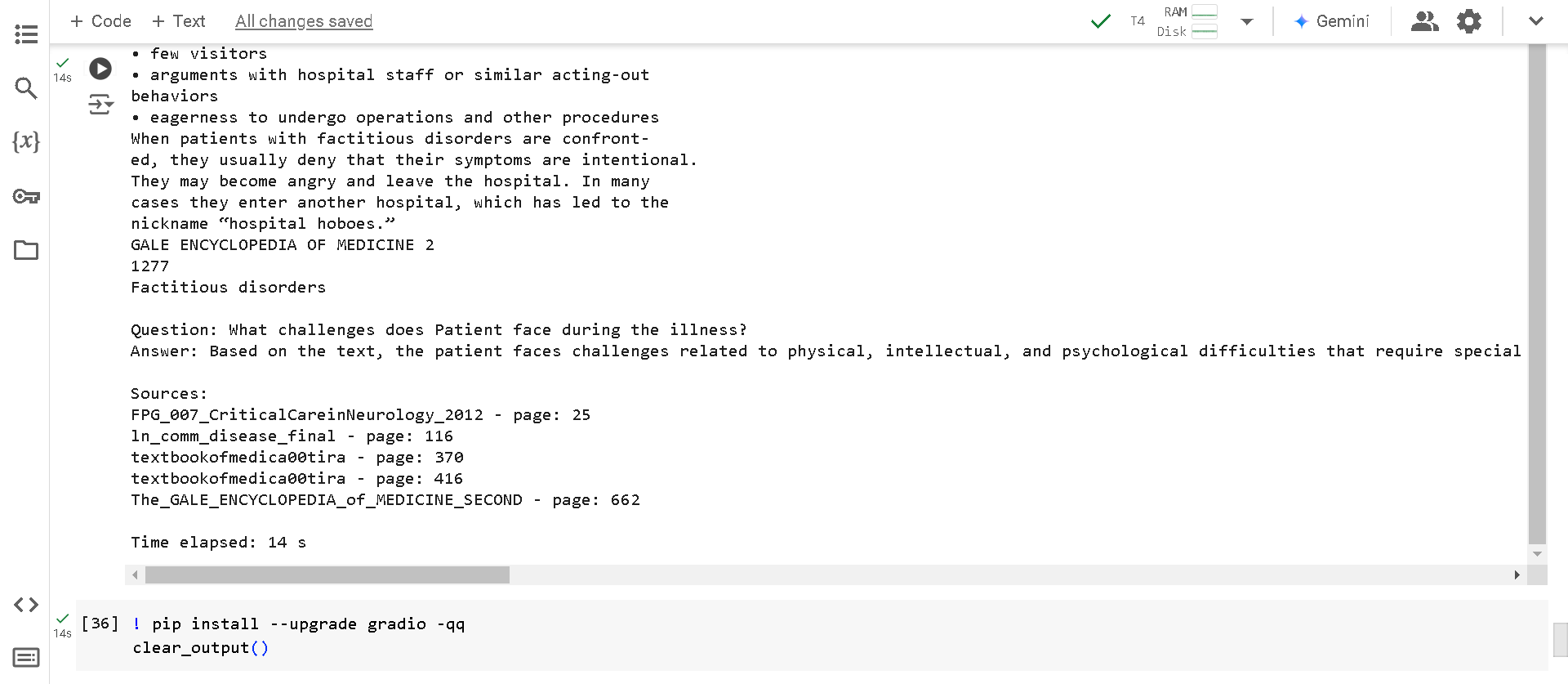


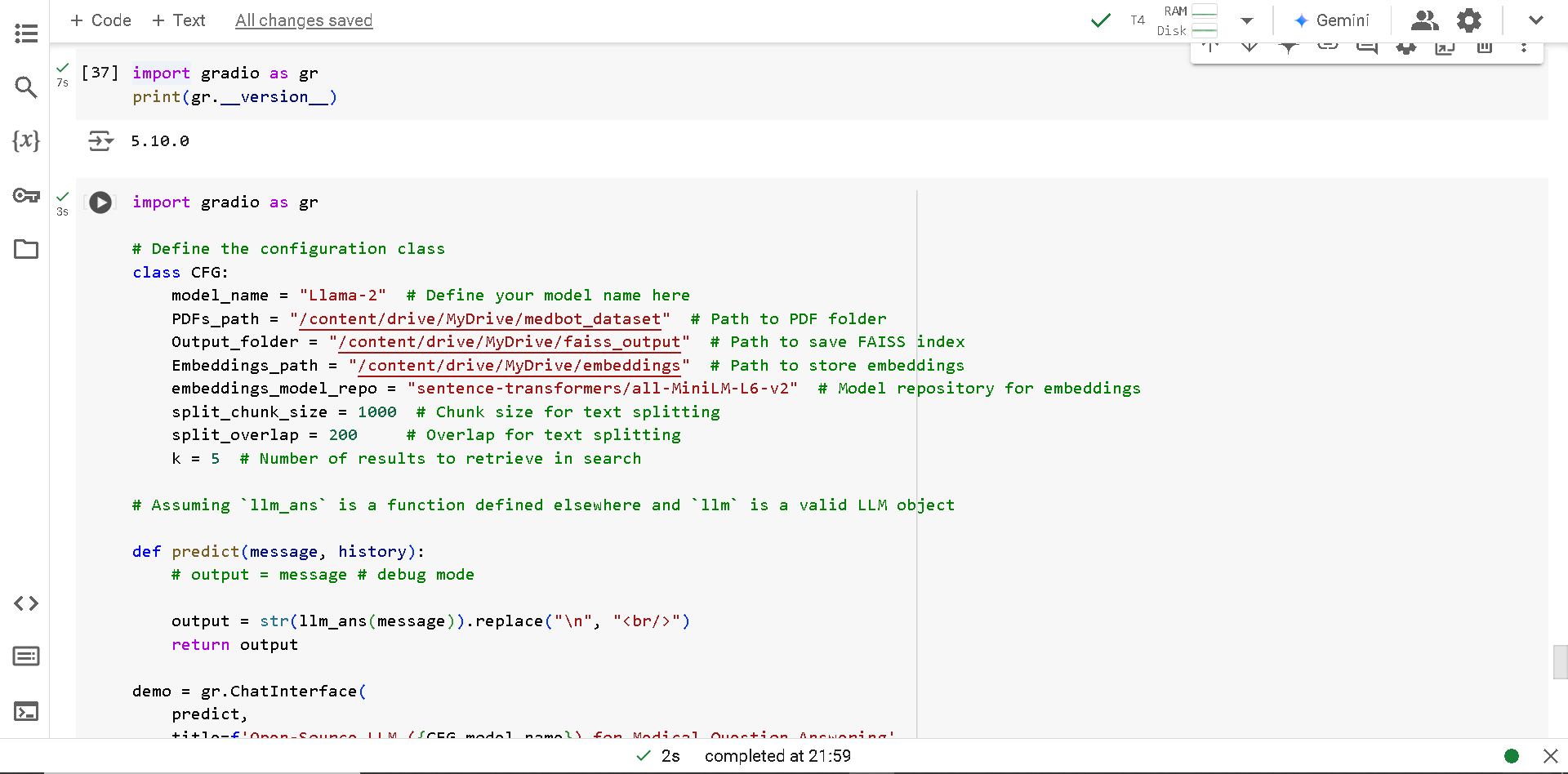


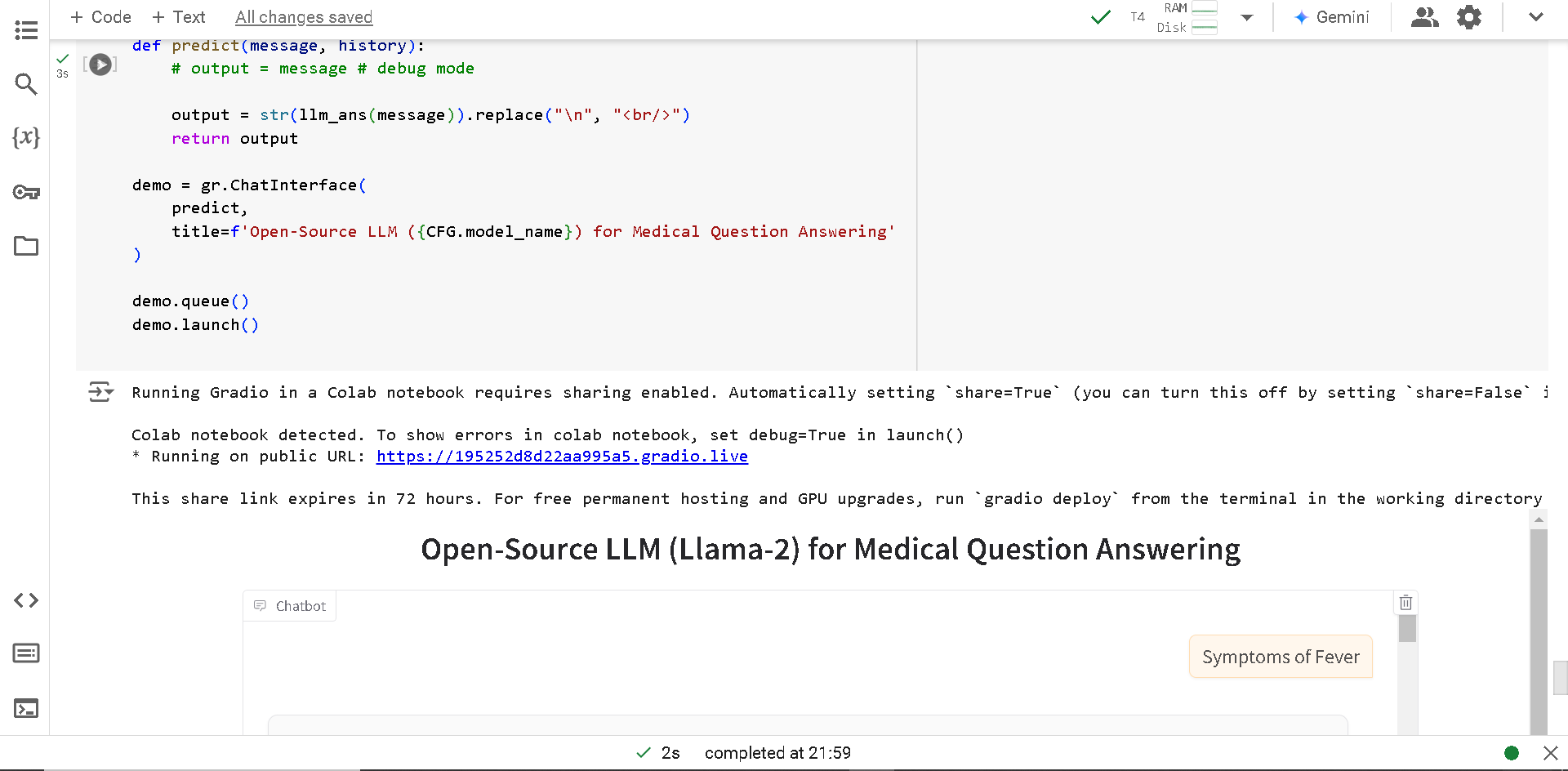




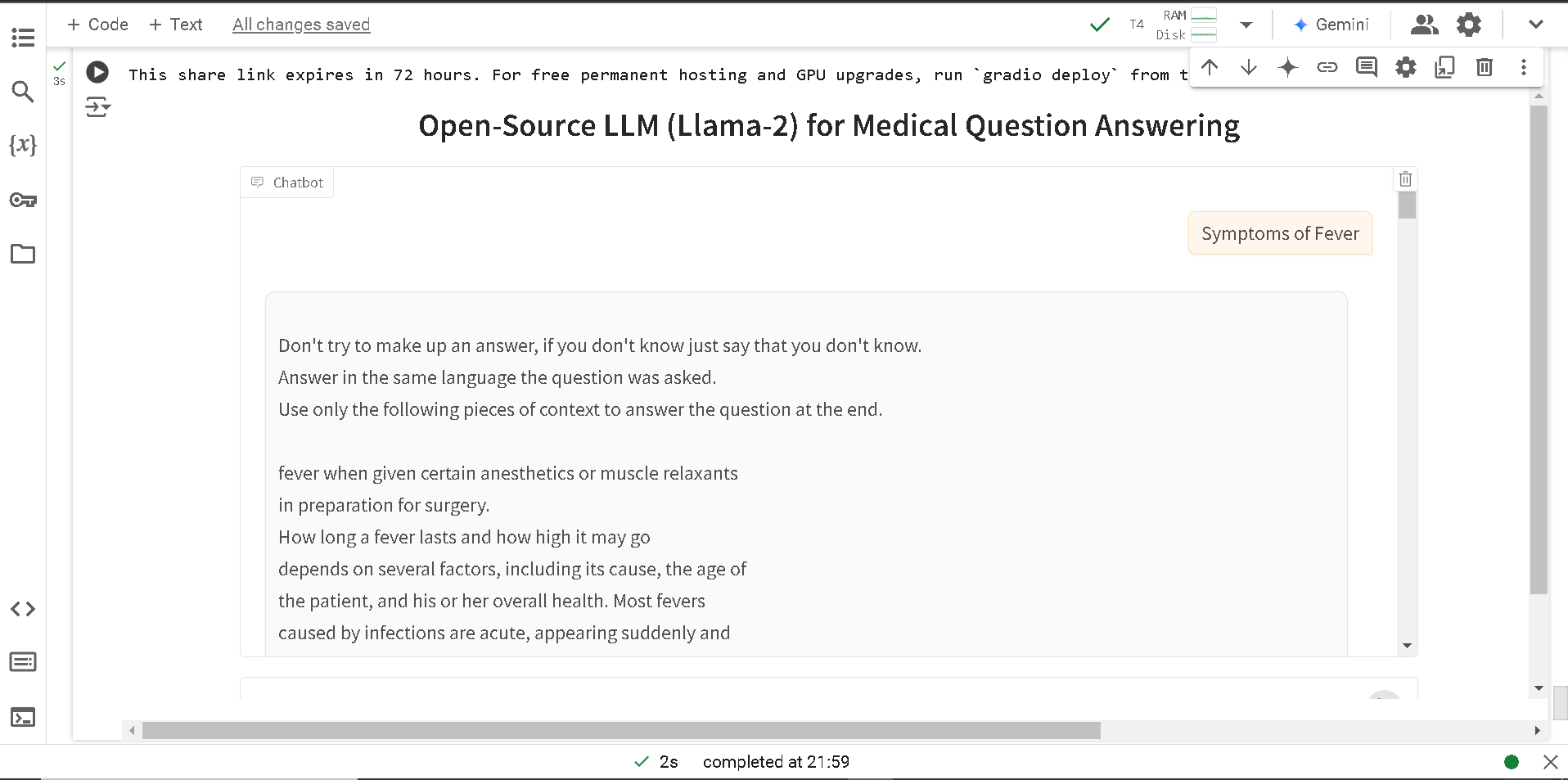


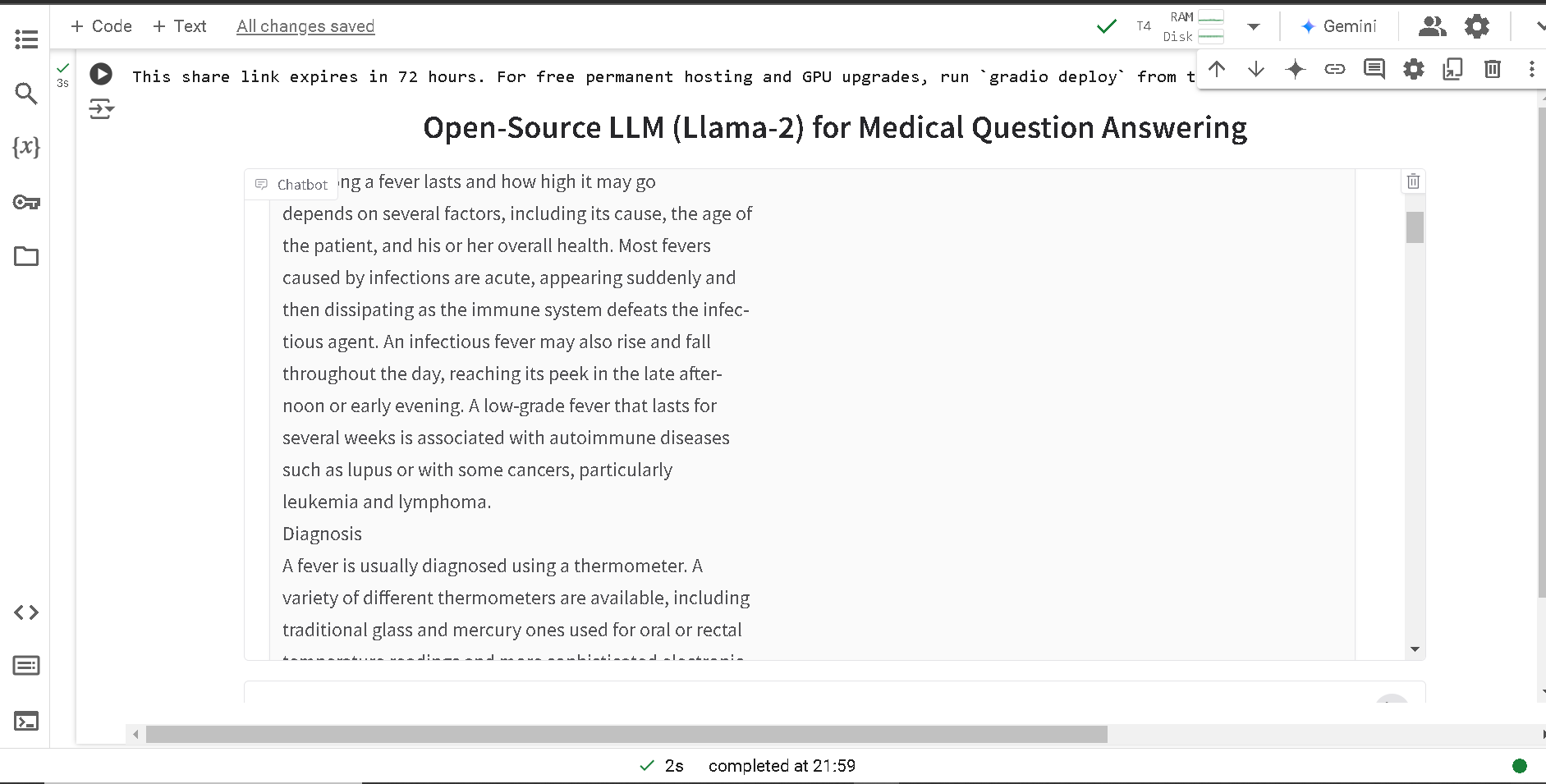






**Output\_screen:**





**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**