# DocuBot: Using Artificial Intelligence To Answer Questions Based on Documents

Kenny Nguyen
University of Waterloo

Pavitar Saini
University of Waterloo

Sami Yousef
University of Waterloo

*Date: Apr. 5, 2023*

**Abstract**

DocuBot is a revolutionary study tool that utilizes AI technology to provide quick and accurate answers to questions based on the context of user-uploaded documents. DocuBot employs several steps including Contextual Understanding, Input Vectorization, and Document Retrieval to successfully answer questions based on user-uploaded documents. In this white paper, we will explore how DocuBot works under the hood and address some frequently asked questions.

## 1 Introduction

In recent years, AI technology has rapidly advanced, leading to the development of innovative tools that can transform the way we learn and work. DocuBot is one such tool that leverages AI technology to provide quick and accurate answers to questions based on the context of user-uploaded documents. By employing advanced algorithms such as Contextual Understanding, Input Vectorization, and Document Retrieval, DocuBot can effectively analyze and understand complex documents, allowing users to easily retrieve relevant information. In this white paper, we will explore the inner workings of DocuBot and address some common questions surrounding this revolutionary study tool.

## 2 Under the hood

### 2.1 Contextual Understanding

DocuBot uses a combination of Optical Character Recognition (OCR) and Natural Language Processing (NLP) to understand the context of the document. Documents are chunked into 25-page chunks and processed through a proprietary OCR algorithm to be converted to LaTeX. This allows DocuBot to understand both natural languages as well as mathematical symbols and formulae.

### 2.2 Vectorizing Input

After the document is processed, DocuBot vectorizes the document using a combination of word embeddings and sentence embeddings. Vectorized documents are stored in Weaviate[4], a vector database that uses the Rust Programming Language to store vector data quickly and safely. The documents are also indexed by Weaviate to provide faster searches are retrievals. **It's important to note that Weaviate can be swapped for ant vector database. We chose Weaviate for its speed and self-hosted solution.**

### 2.3 Retrieving Relevant Context

When DocuBot receives a question, it first vectorizes the question using the same method as the documents. It then uses the vectorized question to search the vector database for the most relevant documents. The documents are then ranked by their cosine similarity to the question. The top 5 documents are then passed to a primary Large Language Model (LLM). Currently, DocuBot uses Galactica[3] - a model trained on scientific paper - as the primary LLM. The primary LLM focuses of formulating context for the question, which will be used by the secondary LLM to formulate a response.

### 2.4 Formulating a Response

Once the context has been formulated, the secondary LLM is used to formulate a response. The secondary LLM is a larger model to provide better responses when the question is unrelated to the content. The secondary LLM is instructed to prepend a sentinel value to the output if the question is not related to the context. The output is then processed to remove the sentinel value and add a message informing the user that the context was not satisfactory to answering the question.

For example, if a user uploads a business textbook and asks "How old is the sun?", the output of the primary model might be:

"NO_CONTEXT The sun is 4.6 billion years old."

which can be processed to:

"That information was not found in your documents,
but I think the sun is 4.6 billion years old."

Currently, DocuBot uses OpenAI's gpt-3.5-turbo model as the secondary LLM, which can be changed for any other LLM.[1]

# 3 Usage

## 3.1 Online Application

DocuBot provides a web application that allows users to upload documents and ask questions. It has an intuitive and user-friendly UI similar to chat applications like ChatGPT.

DocuBot is currently available as a web application at `https://docubot.samiyousef.ca`. However, due to the cost of hosting a server to run DocuBot, this application will soon be deactivated.

## 3.2 Self Hosting

DocuBot can be self-hosted on any machine with sufficient resources. The only requirement is that the machine can run the LLMs used by DocuBot. Depending on the resources on the machine, the LLM can be replaced to yield better or worse results. Additionally, DocuBot can use Nvidia's CUDA or Apple's Neural Engine for hardware acceleration, making it feasible to run on most modern devices.

# 4 Extensiblity and Scaling

DocuBot is designed to be easily extensible and scalable. The LLMs used by DocuBot can be replaced with any other LLM, allowing users to choose the best model for their use case. Additionally, DocuBot can be scaled to support a higher load by simply replicating each instance of DocuBot and adding a load balancer. This is possible because DocuBot is stateless by design.

# 5 Related Work

There are many generative question-answering systems, but we could not find any that are specifically designed to answer questions based on user-uploaded documents. Lilian Weng[5] of OpenAI uses a retriever and reader approach similar to DocuBot.

The Multi-Modal Machine Comprehension (M3C)[2] task aims at answering multimodal questions given a context of text, diagrams and images. The authors found that current models struggle with this task and introduce a dataset which opens new challenges in question-answering systems. In the future, we hope to add support for images, diagrams, and charts; and score DocuBot on the M3C task.

# 6 FAQ

1). *What file formats does DocuBot support?*

Due to time constraints, DocuBot currently only supports PDF files. However, DocuBot can be easily

extended to support other file formats. Other file formats can be processed in the same way as PDF files.

2). *Can DocuBot provide answers to any question?*

Yes, DocuBot will use the secondary LLM to provide an answer to any question not contained in the user-uploaded documents. While the response is usually accurate, it is not connected to the internet and can occasionally produce incorrect answers. It has limited knowledge of world and events after 2021 and may also occasionally produce biased content.

3). *Is DocuBot suitable for use by both students and professionals?*

Yes, DocuBot is suitable for use by both students and professionals. However, DocuBot is not suitable for use in high-stakes situations such as job interviews or exams. DocuBot is not a replacement for a human tutor or teacher, and it should not be used as such. DocuBot is designed to be used as a study tool to supplement learning.

# Bibliography


[1] Tom B. Brown et al. *Language Models are Few-Shot Learners*. 2020. arXiv: 2005.14165 [cs.CL].

[2] Aniruddha Kembhavi et al. "Are You Smarter Than a Sixth Grader? Textbook Question Answering for Multimodal Machine Comprehension". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 5376–5384. DOI: 10.1109/CVPR.2017.571.

[3] Ross Taylor et al. *Galactica: A Large Language Model for Science*. 2022. arXiv: 2211.09085 [cs.CL].

[4] *Weaviate - Vector Database*. URL: https://weaviate.io/.

[5] Lilian Weng. *How to Build an Open-Domain Question Answering System*. Oct. 2020. URL: https://lilianweng.github.io/posts/2020-10-29-odqa/.