

# RAG

## Retrieval-Augmented Generation



גלעד מרקמן  
פרויקט ב- Github

# מה זה RAG

- RAG - Retrieval-Augmented Generation – אלגוריתם שמאפשר לנו לשלב מודל בינה מלאכותית עם מאגר ידע חיצוני.
- מודלי בינה מלאכותית מוגבלים במידע שברשותם. אין להם מידע פרטי והמידע שברשותם אינו עדכני.
- יתרונות שימוש ב RAG:
  - מאשר שימוש במידע עדכני או מורחב
  - נותן תשובות מבוססות על מקור מידע שנקבע על ידינו (ולא על זיכרון המודל).
  - מפחית הזיות (Hallucinations) של המודל.
- מתאים ליישומים כגון:
  - שאלות על מסמכים משפטיים.
  - ניתוח מידע רפואי.
  - חיפוש ידע עסקי פנימי.



# האלגוריתם של RAG

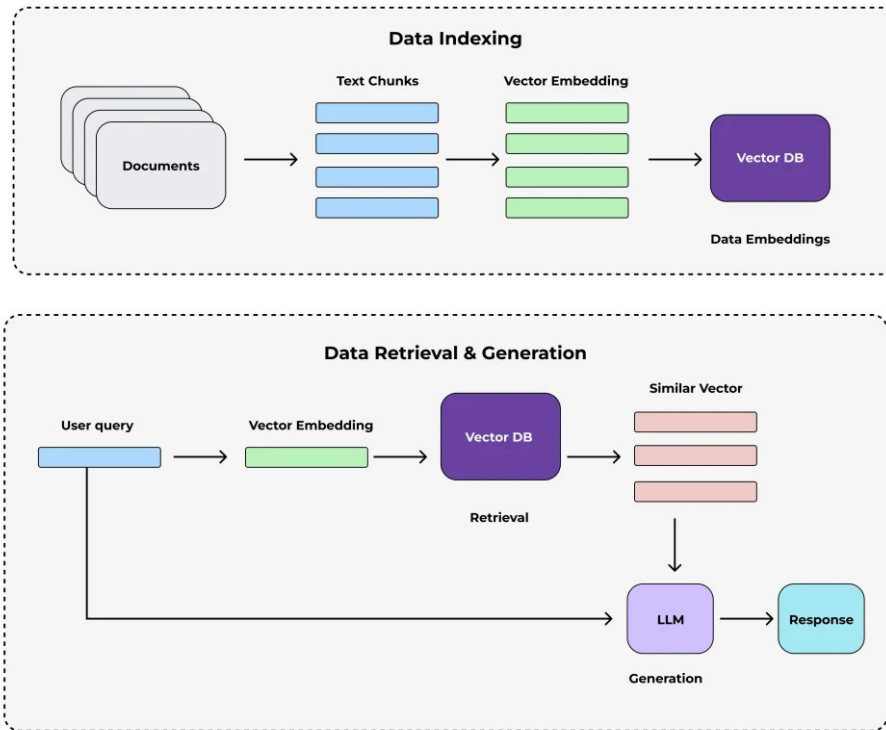
• האלגוריתם RAG בנוי משלושה חלקים עיקריים:

• Data Indexing – המרת הטקסט לזוקטורים (Embedding) ושמירת הטקסט והזוקטורים במסד נתונים מתאים.

• Data Retrieval – שליפת הזוקטורים הרלוונטיים לשאלתה של המשתמש ממסד הנתונים הזוקטורי באמצעות Cosine Similarity.

• Response Generation – יצירת התשובה של מודל הבינה המלאכותית לשאלתה של המשתמש, על סמך הטקסט שנשלף מבסיס המידע.

A DEPICTION OF HOW THE RAG PIPELINE OPERATES



# ChromaDB

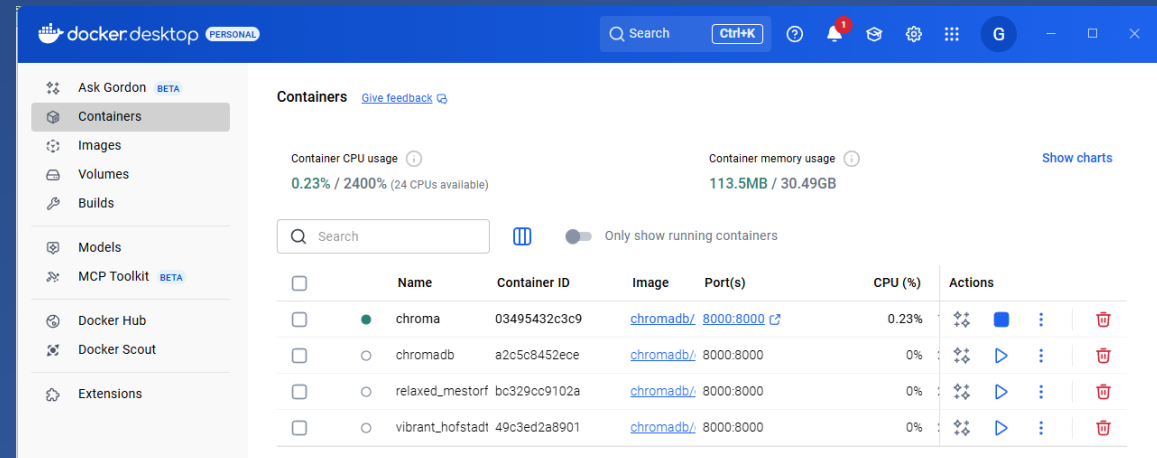
- ChromaDB – הינו מסד נתונים לווקטורים (Vector Database).
- מסד הנתונים מתמחה באחסון והשוואה של Embeddings, וכולל:
  - חיפוש לפי Cosine Similarity המחזיר לנו K תשובות הקרובות ביותר לשאילתה.
  - תמיכה במטה דאטה (פרט לווקטור והטקסט אנחנו יכולים לשמור מידע נוסף, כגון שם המסמך, עמוד, תאריך וכל מידע נוסף אחר שנבחר).
  - תמיכה ב Collections להפרדה של מידע.
  - למסד הנתונים יש שרת נוח הניתן להתקנה מקומית במחשב.
- ב־SK נוספה מחלקה ייעודית המאפשרת עבודה נוחה ופשוטה מול מסד הנתונים באמצעות API מובנה.
- מסד הנתונים בנוי מקוד פתוח וחינמי.



# התקנת ChromaDB

- ChromaDB הינו שרת בסיס נתונים הרץ בענן או על המחשב שלנו, ואלינו  
אנו מבצעים קריאות API לאחסון או שליפת נתונים.
- על מנת להריץ ChromaDB על המחשב נבצע את הפעולות הבאות:
  - גשו לאתר הרשמי של Docker: <https://docs.docker.com/>
  - התקינו Docker Desktop לפי מערכת ההפעלה שלכם.
  - Visual Studio פתחו Terminal (Developer PowerShell) והריצו:  
`docker run --name chroma -p 8000:8000 -d chromadb/chroma:0.4.24`

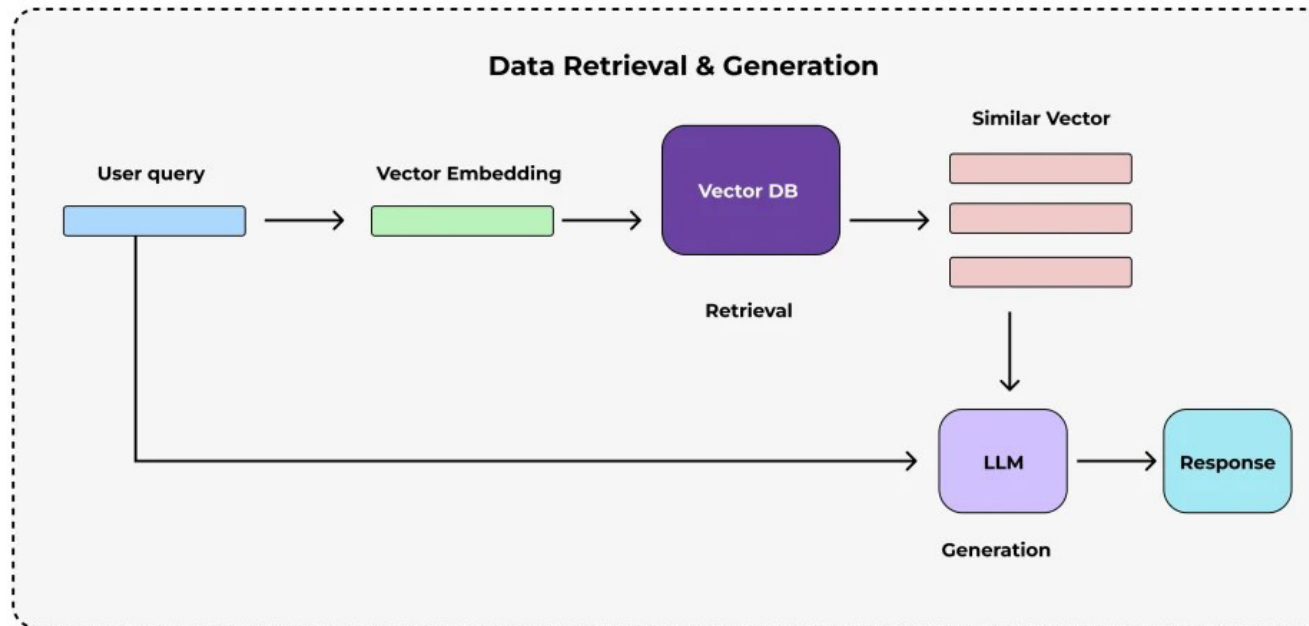
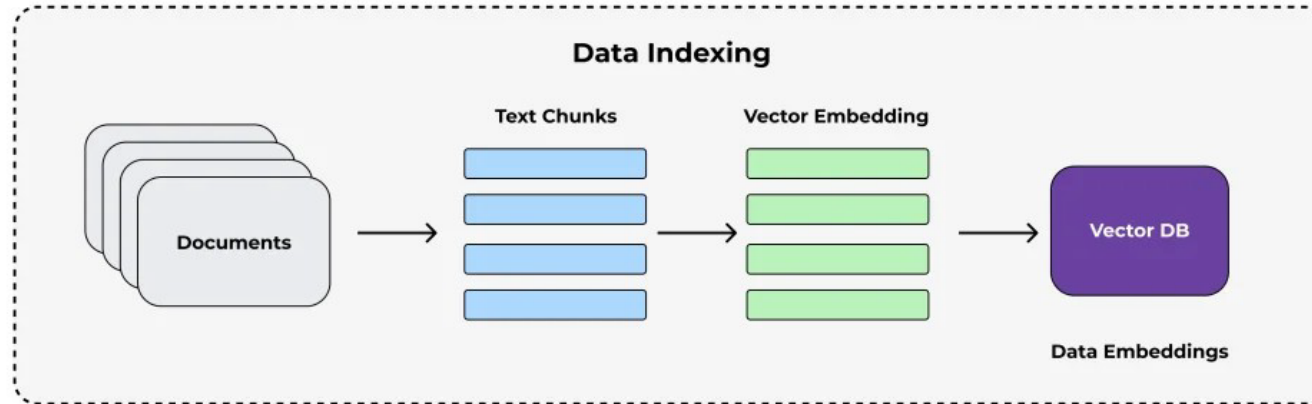
```
Developer PowerShell
+ Developer PowerShell | [Icon] [Icon] [Icon]
*****
** Visual Studio 2022 Developer PowerShell v17.14.16
** Copyright (c) 2025 Microsoft Corporation
*****
PS C:\Users\Gilad\source\repos\SK> docker run --name chroma -p 8000:8000 -d chromadb/chroma:0.4.24
```



# מחלקות הפרויקט שלנו

- Program.cs – התוכנית הראשית אשר מאפשרת טעינת מסמך PDF, המרתו לזוקטורים ושמירתם ב ChromaDB. הפעלת השיחה עם המודל על בסיס המידע הזוקטורי.
- PdfLoader – מחלקה לטעינה של מסמך PDF והמרתם לאצוות של טקסט. המחלקה משתמשת בספירה iText.
- ChromaClient – מחלקה המטפלת בהוספת מסמכים לבסיס המידע הזוקטורי, חיפוש מסמכים, ומחיקת Collection של מסמכים. המחלקה מטפלת בשלב ה Indexing.
- RagChat – מחלקה מטפלת בשלב ה Data Retrieval + Response Generation.

## A DEPICTION OF HOW THE RAG PIPELINE OPERATES



# RAG Pipeline