אוניברסיטת אריאל בשומרון

# LLM Evaluator

| ת.ז | שם מלא |
|---|---|
| 209349356 | שייה טל אורנן |

## טכנולגיות בשימוש:

| ממשק משתמש: | שכבת נתונים: | שפות תכנות: | טעינת מודלים של LLM | UI |
|---|---|---|---|---|
| REST API | MySQL | Python | הספריה: Hugging Face Transformers | הספריה : Streamlit |
| FastAPI | | | | CSS |
| | | | | HTML |

## מטרת הפרויקט (מערכת מידע):

מערכת להערכת תשובות LLM. מערכת זו נועדה לספק כלי רב-תכליתי להערכת איכותן של תשובות המופקות על ידי מודלים גדולים של שפה (LLM).

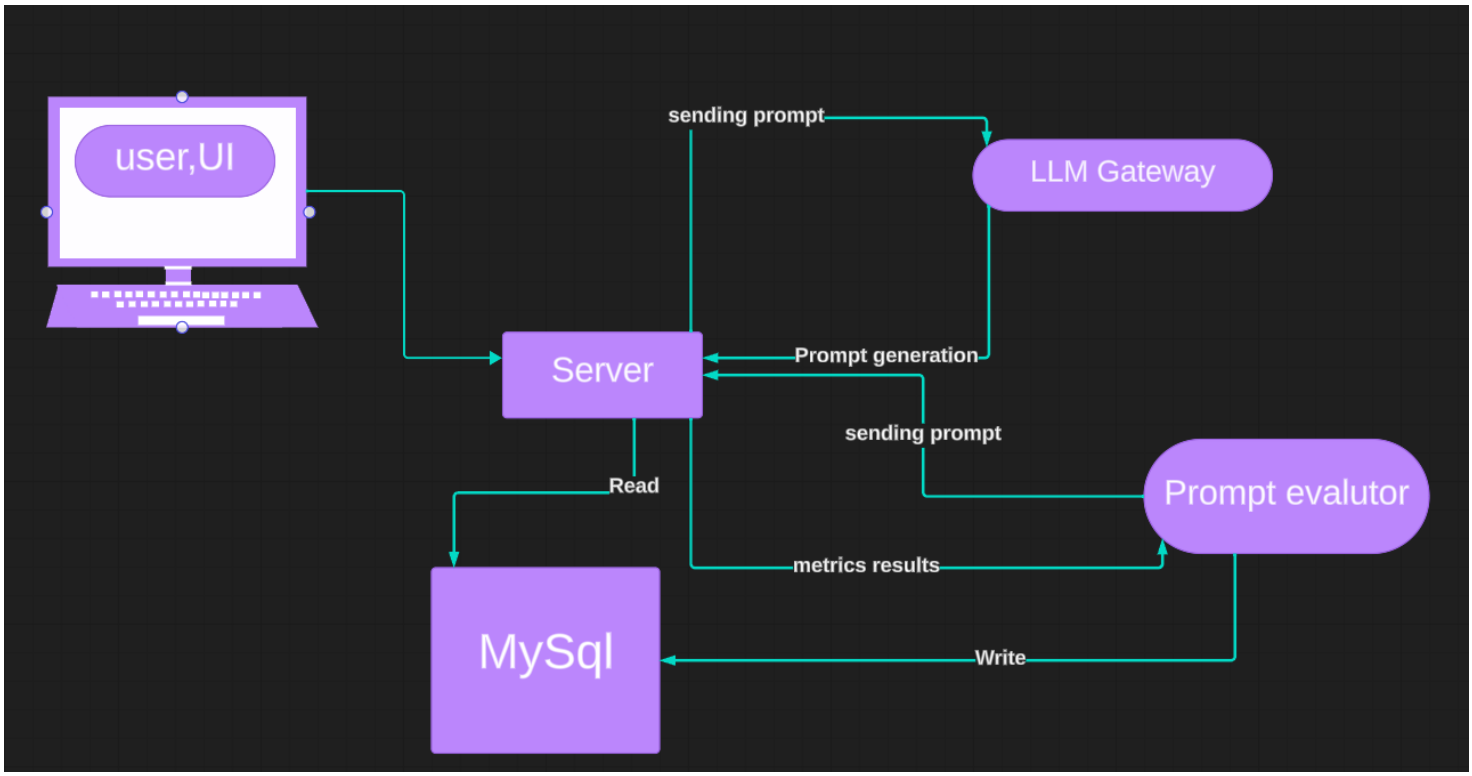
**פונקציונליות:**

- הגדרת Prompt (שאלה או הנחיה)
- בחירת LLM (מודל רצוי) מתוך רשימה זמינה
- בחירת מטריקות להערכה, כגון Rough score, Coherence, Fluency, Toxicity
- ניתוח תוצאות ותשובות המודל
- השוואת ביצועים של מודלים שונים עבור אותה משימה
- ניתוח נתונים לאורך זמן וזיהוי מגמות

# הרשאות

| | |
|---|---|
| מהנדס פרומפטים | אחראים על יצירת Prompts עבור llm.<br>מעוניינים להעריך את איכות ה-Prompts שלהם.<br>רוצים להשוות בין Prompts שונים.<br>רוצים לקבל סטטיסטיקות על ביצועי ה-Prompts שלהם. |
| מפתח מודל/אנליסט | אחראים על פיתוח מודלים עבור llm.<br>מעוניינים להעריך את ביצועי המודלים שלהם.<br>רוצים להשוות בין מודלים שונים.<br>רוצים לקבל סטטיסטיקות אגרגטיביות על ביצועי המודלים שלהם. |
| | |

# ארכיטקטורת המערכת:




# תיאור קצר של המערכת המוצעת

llm הוא כלי רב עוצמה ליצירת טקסט, תרגום שפות, כתיבה יצירתית ועוד. עם זאת, קשה להעריך את איכות התשובות של llm באופן ידני. מערכת זו נועדה לפתור את הבעיה הזו על ידי:

הערכת תשובות llm באופן אוטומטי: המערכת תשתמש במגוון מטריקות טקסטואליות (כגון ROUGE, קוהירנס וכו') כדי להעריך את איכות התשובות של llm.

**<u>השוואה בין תשובות של מודלים שונים:</u>** המערכת תאפשר למשתמשים להשוות בין תשובות של מודלים שונים עבור Prompt נתון.

**<u>יצירת סטטיסטיקות אגרגטיביות עבור מודלים:</u>** המערכת תיצור סטטיסטיקות אגרגטיביות עבור מודלים, כגון:

דיוק ממוצע, זמן תגובה ממוצע ועוד.

אפשרות לשיתוף תוצאות עם משתמשים אחרים: המערכת תאפשר למשתמשים לשתף את תוצאות הניתוח שלהם עם משתמשים אחרים.

**<u>צורך במערכת:</u>**

הערכת תשובות llm באופן ידני היא קשה ודורשת וזמן רב: llm יכול ליצור תשובות ארוכות ומורכבות, וקשה להעריך את איכותן באופן ידני.

# תהליכים משמעותיים לכל סוג הרשאה (תהליך אקטיבי, ולא צפייה בנתונים)

**הרשאה 1: מהנדס פורמטים**

1. יצירת פרופיל: בחירת מודלים רלוונטיים, שם משתמש וסיסמה.
2. יצירת Prompt: הזנת Prompt, בחירת מודלים להערכה ובחירת מטריקות.
3. צפייה בניתוח מטריקות של הprompt, ומטריקות אגרטביות עבור היוזר הספציפי.
4. המלצות לשיפור הprompt

**הרשאה 2: מפתח מודל**

1. הרשמה: בקשת הרשאות עבור מודל/מודלים.
2. ניתוח ביצועים: צפייה בניתוח הביצועים של כל מודל עבור מטריקות ספציפיות.
3. מסך המציג המלצות באיזה מודל מומלץ להשתמש עבור prompt ספציפי בהתאם לנתוני עבר.

# <u>מסכי</u> האפליקציה (מסך כניסה+ תהליך משמעותי של כל הרשאה + פרופיל משתמש לכל הרשאה):

מסך 1-מסך בית:

מסך 2-5 - מסכי הרשמה - במידה ואין משתמש- יוצרים משתמש חדש.

ובמידה ויש -התחברות

מסך 7-6-תהליך משמעותי עבור מפתח מודל/אנליסט:צפייה בניתוח הביצועים של כל מודל עבור מטריקות ספציפיות.

מסך 8+7 - תהליך משמעותי עבור מהנדס פרומפטים:הזנת Prompt, בחירת מודלים להערכה ובחירת מטריקות.