דו"ח תרגיל בית 5

Transformer and LLMs

שמות המגישים + ת"ז:

עוביידה חטיב, 201278066

מאיה עטואן, 314813494

חלק א'

הפונקציה data_load, שנקראת מה- main, אחראית על טעינת ה- subset מתוך המאגר. אם יש subset שנשמר מקודם, היא טוענת אותו מהדיסק, הדבר נעשה ע"י בדיקת אם התיקייה של הפרויקט מכילה מאגר נתונים בשם 'imdb_subset' . אם כן, היא טוענת אותו. אחרת, הפונקציה טוענת את מאגר הנתונים השלם של IMDB, בוחרת באופן אקראי 500 דגימות ממנו, ושומרת אותו לדיסק. לאחר שמירתו לדיסק, היא טוענת אותו מחדש.

חלק ב'

המודל bert-base-uncased נטען, והוגדר לביצוע סיווג לשתי קטגוריות (ע"י num_labels=2) בהתאם לסיווג שעושה המשימה שלנו לביקורות לחיוביות ושליליות. המודל עבר טוקנזציה, ושם העמודה 'label' הוחלף ל- 'labels'. מאגר הנתונים שלנו חולק לשני סטים: אימון (80%) ובדיקה (20%). לאחר מכן, אובייקט Trainer אותחל, כאשר קיבל לצד המודל את הארגומנטים של האימון, שני הסטים של האימון והבדיקה, הטוקנייזר, והפונקציה compute_metrics המחשבת את דיוק המודל. הוא עבר אימון ואיבלואציה, כאשר את תוצאות האיבולואציה מודפסות למסך. **התוצאה מצורפות להלן**:

```
0.93
```

האוגרמטים שנבחרו עבור האימון:

evaluation_strategy = "no", נבחר בשל כך שאנחנו לא נדרשים להדפסה במהלך האימון, אלא רק את התוצאות שהתקבלו בשלב האיבולאציה.

learning_rate= 1.5e-5, זהו הערך שהציג איזון בין יציבות המודל ומהירות הלמידה שלו. במיוחד, הוא הראה תוצאות יותר טובות מאלה שהערך הסטנדרטי 2e-5 הראה, כנראה בשל היות מאגר הדוגמאות שלנו קטן יחסית.

per_device_train_batch_size = 8, הערך התאים למאפיינים של המחשב שלנו ולכן לא היה צורך בלשנות אותו.

num_train_epochs = 3 , הערך נבחר על סמך תצפית שלנו לתוצאת שהראו שמעבר ל- epoch השלישי אחוז הדיוק וה- function loss לא משתנות בהרבה, ומזה הוסק ששימוש במספר גדול יותר עלול להוביל ל- overfitting.

weight_decay = 0.005, גם כן נבחר על סמך תצפית שלנו, כך שהערך הראה איזון והתאמה למשתנה num_train_epochs.

הפונקציה compute_metrics שהוזכרה קודם לכן מחשבת את הביצוע של המודל ע"י השוואת הפלט logits לתוויות האמיתיות labels. היא משתמשת בפונקציה np.argmax() על מנת להמיר את התוויות של הפלט לבינאריות (0,1) ומשווה אותם לתוויות האמיתיות ע"י הפונקציה accuracy_score() מהספרייה sklearn.

שאלות

1. **האם התוצאות היו טובות? האם הן תאמו לציפיות שלכם? הסבירו.**

לדעתנו, התוצאות טובות, התוצאה של 93% מצביעה על דיוק גבוה. הם כן תאמו במידה רבה לציפיות שלנו, בשל כך שהמודל יותר דינמי בתפיסת המבנה התחבירי של מילה במשפט, ואת ההקשר הסמנטי בין המלים לעומת מודלים שהשתמשנו בהם בעבר כמו TF-IDF.

2. **אם נרצה לסווג ביקורות ספרים באמצעות אותו מודל שאימנו, האם לדעתכם התוצאות יהיו טובות? הסבירו.**

אנחנו מצפים שהמודל יצליח להתמודד עם סיווג ביקורות ספרים, אולם במידה פחות טובה. יש הרבה משותף בין שתי המשימות. דוגמה לכך היא השימוש בשם תואר על מנת לתאר את החוויה או היבטים בה. שם התואר תמיד יצביע על חוויה שהיא כפן שלו, כמו במקרה של המילה "מאכזב" שנוטה להופיע בחוות דעת שלילית יותר מאשר חיובית ללא קשר להקשר. מצד שני, יש אלמנטים שכן ייחודיים להקשר של סרטים, כמו האלמנטים הוויזואליים הקשורים לצילום והעריכה, או המשפט שהוזכר בהרצאה " the

value I got from the two hours watching it was the sum total of the popcorn and the drink" שיש לה פחות משמעות כאשר היא מופיעה בביקורת של ספר.

3. **אם נרצה לסווג ביקורות סרטים ממאגר אחר, שאינו IMDB , באמצעות המודל שאימנו, האם לדעתכם התוצאות יהיו טובות? הסבירו.**

אנחנו מצפים שהתוצאות יהיו יותר טובות מאלה של הספרים, אולם פחות בהשוואה לביקורות של מאגר ה- IMDB, כאשר מידת הדיוק תהיה בהתאם לכמה המאגר הזה דומה למאגר של ה- IMDB. אלמנטים שיכולים להיות שונים בין שני המאגרים: סגנון השפה, למשל אם הביקורות לקוחות ממגזין שבו אנשים מקצועניים כותבים את הביקורות אנחנו נצפה לשפה יותר פורמלית ושימוש באוצר מלים שונה. ז'אנרי הסרטים שהוא מכיל והתפלגות הז'אנרים השונים מתוך כלל הסרטים, כי ז'אנרים שונים מושכים קהלים שונים בתחביב, בגיל וכתוצאה מכך המלים, האורך של הביקורת ועוד מאפיינים עשויים להיות מושפעים.

4. **עבור כל אחד מהמקרים מסעיפים 2,3 , תארו מה נוכל לעשות כדי לשפר (עוד יותר) את התוצאות?**

להן הצעות שאמורות לשפר את התוצאות של המודל. ההצעות תקפות לשתי המשימות מהסעיפים 2,3:

- להוסיף דאטה שהיא ביקורות ספרים לסט האימון (להוסיף דאטה ממאגר הסרטים האחר), ובכך ליצור מודל המסוגל לסווג את שני הסוגים של ביקורות.
- לעשות fine tuning למודל ולהתאים אותו מחדש למשימה של סיווג ספרים (להתאים אותו לסגנון ותוכן הביקורות של המאגר האחר).
- להגדיל את הרגולריזציה, ע"י לתת ערך גדול יותר לארגומנט weight_decay אשר מונע למידת יתר של סגנון מסוים, דבר שאמור להשתקף ביכולת המודל להכליל טוב יותר למשימה של ביקורות הספרים (להכליל טוב יותר לתבניות אחרות של ביקורות סרטים).

חלק ג'

הקוד מתחיל בטעינת מאגר הנתונים של IMDB, ומתוכו הוא בוחר באקראי 100 ביקורות חיוביות ו- 100 שליליות. הקוד לאחר מכן טוען שני מודלים נפרדים של GPT-2, אחד ישמש לאימון על ביקורות חיוביות והשני על ביקורות שליליות. כמו כן, לכל מודל, טוענים גם את הטוקנייזר שלו. השלב שאחרי הוא ביצוע טוקינזציה על המאגרים שבחרנו. הפעולה

מתבצעת ע"י הקריאה לפונקציה tokenize_reviews(), אשר ממירה את הביקורות לרצפי טוקנים קבועי אורך. התהליך מתבצע עבור כל סוג של ביקורות (חיוביות ושליליות) בנפרד.

בשלב שאחרי, נוצרו שני מופעים של TrainingArguments, שהגדירו את פרטי האימון של המודל. **בחרנו את הערכים של הפרמטרים** באופן זהה לרוב לערכים שהשתמשנו בהם עבור המשימה בחלק ב' מטעמים דומים: evaluation_strategy = "no" כי לא התבקשנו להדפיס בשלב האימון, ה- learning_rate= 2e-5 הוא הערך הכי בשימוש, ובשל היות המאגר שלנו לא גדול במיוחד העלאתו יכולה להביא ל- Overfitting.
per_device_train_batch_size = 8, התאים למאפיינים של המחשב שלנו. באשר ל- num_train_epochs ו- weight_decay נבחרו להיות 3 ו- 0.01 בהתאמה על סמך תצפית שערכנו, ערכים אלו הראו איזון טוב בין Overfitting ל- Underfitting. יש לציין שהערכים של הפרמטרים זהים עבור שני המודלים בשל החשיבות שהתהליך יתבצע באופן אחיד עבור שני סוגי הביקורת.

לאחר מכן, נוצרו שני אובייקטי Trainer, שכל אחד שימש לאימון אחד המודלים. המודלים והטוקנייזר נשמרו לשימוש עתידי, ונטענו שוב מהדיסק. בשלב שאחרי, הוגדר prompt להיות "The movie was ", הוא עבר קידוד והתווסף אליו attention mask. הפעולה התבצעה בנפרד עבור כל מודל.

לאחר מכן, הפונקציה reviews_generate נקראה פעמיים עבור כל סוג של ביקורת. הפונקציה מפיקה (generate) חמשה ביקורות ע"י השימוש בפונקציה generate() של המודל עם **הפרמטרים והערכים הבאים**:

- **max_length = 100**, הערך של הפרמטר נבחר להיות כזה משום שערכים גבוהים יותר הראו נטייה של הביקורת לכלול ציטוטים יותר מאשר שהן מביעות דעה. הציטוטים מובאים כדברים שנאמרו בתוך הסרט ע"י אחד הדמויות. מנגד, ערך max_length קטן יותר הציג ביקורות שטחיות שלא מביעות דעה מעמיקה כך שלפעמים קשה לסווג אותה כחיובית או שלילית גם עבור מישהו אנושי.
- **do_sample = True**, מאפשר רנדומליות בבחירת המילה בכל שלב שלא בהכרח המילה בעלת ההסתברות הגדולה ביותר. זה חשוב משום שהמשימה שלנו היא הפקת משפטים בעלי משמעות וחשוב לתת מקום ליצירתיות ולא משימה דטרמיניסטית שבה יש מילה הנכונה בהכרח. כמו כן, זה מתבסס על תרגילי הבית הקודמים שראינו בהם שיש לפעמים השלמות שהן פחות שכיחות מהשלמות אחרות, כלומר שהמילה המשלימה לא בהכרח בעלת ההסתברות הכי גדולה.

- **temperature=0.5, top_p=0.9**, שני הפרמטרים קשורים אחד לשני, וקובעים את מידת אקראיות\שמרנות המודל. הערך נבחר על סמך התצפיות שערכנו, כאשר ערכים גבוהים יותר אולם יצרו משפטים נכונים תחבירית, אך לא היו נכונים מבחינה עובדתית. דוגמה לכך " Anne Rice and her husband's … dog Harry Potter when he met William Shakespeare …". כמו כן, ערך נמוך יותר ל- temperature הופך אותו לדטרמיניסטי, ואילו ערך נמוך של top_p מתעלם מהרבה השלמות שהן לא שכיחות במיוחד, למשל: למשל המילה "הקניק" במשפט " אגם **הנקיק** הוא יעד טבעי יפהפה המוקף בנוף עוצר נשימה".
- **repetition_penalty=1.2**, הבחירה נועדה למנוע מהמודל לחזור על מלים באופן מיותר ובכך לשפר את הנראות של הטקסט. לפעמים, אכן אין ברירה אלא להשתמש באותה מילה, למשל, במקרה של שם פרטי של מישהו, לכן לא נרצה ערך נמוך מדי.
- **num_return_sequences = 5**, זהו מספר המשפטים שהתבקשנו להפיק.

לאחר ההפקה, התוצאות עוברות decode כדי שהתוצאה תהיה טקסט קריא, ונשמרות בקובץ הטקסט generated_reviews.txt לפי הפורמט הנדרש.

שאלות

1. **האם פלטי המודלים תאמו לציפיות שלכם? הסבירו.**

התוצאות שהופקו ע"י המודלים אינן לגמרי מספקות, כי למרות היותן נכונות תחבירית, אינן יוצרות ביקורות אמיתיות ומגובשות, וקשה לזהות דעה ברורה בהם, או סממנים היכולים לעזור בלסווג אותם כחיוביות או שליליות. לפעמים אפילו אין עקביות בין המשפטים.

2. **האם ראיתם הבדלים משמעותיים בתוצרים של כל אחד מהמודלים? פרטו.**

בשל מה שהוזכר בתשובה לשאלה 1, ההבדלים בין שני המודלים אינם חדים מספיק, שכן שניהם אינם מביעים דעה שניתן להגיד עליה בבירור שהיא חיובית או שלילית, אלא שרשור של משפטים שסביר שיופיעו בשני ההקשרים.

3. **הסבירו מה היה משתנה אילו היינו מגדילים ואילו היינו מקטינים באופן משמעותי את כמות הדוגמאות בסטי האימון. התייחסו הן לתוצאות והן לתהליך האימון.**

הגדלת מספר הדוגמאות הייתה מביאה ללמידה יותר טובה של הדפוסים האופייניים לכל אחד משני סוגי הביקורת, כך שהדבר היה משתקף בהבדלים יותר ברורים וחדים בין שני המודלים. מצד שני, ההגדלה של הסט תביא ליותר זמן ומשאבים חישוביים.

הקטנת מספר הדוגמאות, באופן סימטרי, הייתה מביאה ללמידה יותר שטחית, ומגבירה את הרנדומאליות של המודל. מבחינה חישובית, המודל יידרש פחות משאבים וייקח לו פחות זמן אימון.

4. הסבירו מה התפקיד של ה attention mask שיצרתם בסעיף 8.

ה- Attention mask קובע לאיזה חלק מהקלט המודל צריך להתייחס ומאיזה חלק הוא יתעלם. בקוד שלו העברנו את הפרמטר tokenizer.pad_token_id, מה שאומר שנרצה שהמודל יתעלם מטוקני ה- Padding, שהם לא מלים אמיתיות ונמצאות על מנת לשמור על אורך משפט אחיד.

5. עד כמה לדעתכם ה prompt שהעברנו למודל משמעותי עבור התוצאה? התנסו ב prompts אחרים לבחירתכם ובדקו את השערתכם. פרטו בדו"ח.

ה- prompt מאוד משמעותי עבור התוצאה, בשל כך שהוא מהווה את נקודת הפתיחה של הביקורת ומשפיע על הכיוון של ההשלמה שלה. ה- prompt שהעברנו גורם למודל להתמקד בלדבר על סרט, תוך דגש על הבעת דעה.

מפני שלדעתנו התוצאות תוך שימוש ב- prompt הנתון לא היו כמצופה, נסינו שני prompts שהם יותר ממוקדים לעבר הבעת דעה: "This movie made me feel" , שמשתמש במילה "להרגיש" באופן מפורש, ו- "In my opinion, the movie was" המשתמש במילה "דעה" אשר אמורה לכוון את המודל לעבר הבעה.

הניסיון הראשון ("This movie made me feel") הניב תוצאות יותר טובות מה- prompt הנתון לנו והצליח להביא להשלמות שכן מכילים דעה או ביקורת שניתן להסיק ממנה אם הכותב אהב או לא אהב את הסרט. בשונה מהמצופה הרבה פעמים המילה\המלים הראשונות שהשלימו את ה- prompt לא היו שמות תואר או ביטויים רגשיים, אולם בהמשך כן היה התייחסות להתרשמות מהסרט. דוגמה לכך היא המשפט שהתחיל ב- "This movie made me feel like I was in a new country", אך ההמשך שלו היה "and that's the only reason why it doesn't work", מה שנותן לנו להבין שהיחס של הביקורת הוא שלילי. כמו כן, הדעה לרוב כן התיישבה עם הסיווג שלה.

התוצאות של "In my opinion, the movie was" היו עוד יותר טובות מבחינת כך שהם מביעים דעה ומספרים על חוויה והתרשמות, והרבה פעמים היה כבר אפשר להסתפק בלקרוא את המשפט הראשון על מנת להבין את הנטייה של הדעה, כך למשל הפסקה שהתחילה ב- "In my opinion, the movie was really good ...". אולם בשונה מהניסיון הראשון שלנו הדעות לא היו מסווגות באופן נכון תמיד, והיה נראה שהם שובצו באופן אקראי לכל אחת משתי הקטגוריות.

חלק ד'

הקוד מתחיל בטעינת המודל והטוקנייזר. לאחר מכן, הוא בוחר 50 ביקורות מאוזנות (25 חיוביות ו- 25 שליליות) מבין המאגר הנתון כקלט. הדבר נעשה ע"י הקריאה לפונקציה data_load שטוענת את המאגר, מסננת את הדוגמאות שבו לחיובי ושלילי, ובוחרת ב- 25 מבין כל קטגוריה (או פחות במקרה שאין כמות כזאת). לבסוף, הדוגמאות שנבחרו מכל קטגוריה מתאחדות למאגר אחד שמכיל 50 ביקורות.

בשלב הבא 3 פרומפטים נוצרים, אחד מכל סוג, פירוט לגביהם ואת התוצאות שהם הניבו מצורפות למטה.

הסיווג של הדוגמאות נעשה ע"י הקריאה לפונקציה reviews_classify אשר מקבלת כקלט את הביקורות, ביחד עם שלושת הפרומפטים, ומסווגת כל ביקורת אחרי שהיא משרשת אותה עם כל אחד מהפרומפטים, התוצאה נשמרת בתוך מבנה נתונים מסוג מילון. לאחר מכן, הפונקציה save_results() נקראת, והיא שומרת את התוצאה בקובץ הנתון לפי הפורמט הנדרש.

הפרומפטים שבחרנו:

- ה- Zero shot prompt:

Classify the following movie reviews as positive or negative. The review:

- ה- Few shot prompt:

Classify the following movie reviews as 'positive' or 'negative':\n1. 'The movie was absolutely fantastic, I couldn't stop watching!' -> positive\n2. 'It was a complete waste of time, I regret watching it.' -> negative\n3. 'The film had great acting but the plot was weak.' -> negative\n4. 'A must-watch! One of the best movies I've seen in years!' -> positive. The review:

- ה- Instruction-based prompt:

You are a movie reviews classifier: Please classify the following reviews as positive or negative. The review:

ולהלן התוצאות שהם הציגו:

```
Accuracy of zero-shot prompt: 0.9
Accuracy of few-shot prompt: 0.88
Accuracy of instruction-based prompt: 0.9
```

אחוזי הדיוק גבוהים יחסית, ומעידים על כך שהסיווגים של המודל רחוקים מלהיות אקראיים. יש לציין שהבקשה המפורשת, הדוגמאות, וההכוונה בהם השתמשנו עור הפרומפטים Few-shot ו- Instruction based לא השתקפו בשיפור בתוצאות מה שמעיד על אי-היותם גורם משמעותי. עם זאת, יש לקחת בחשבון היות המאגר שלנו קטן, כך שהוא רחוק מלהיות מדגם מייצג.

דוגמאות לסיווגים מוצלחים ולא מוצלחים:

- **Zero shot prompt:**
  - **Truly classified as Positive**: *Not only is this film entertaining, with excellent comedic acting, but also interesting politically. It was made at the end of the Soviet Union, but makes fun of the soviet mentality through and through. The story is set during the early days of the soviet union, and it questions the rationale behind the revolution both in cultural and practical terms...*
  - **Truly classified as Negative**:
    The prerequisite for making such a film is a complete ignorance of Nietzche's work and personality, psychoanalytical techniques and Vienna's history...
  - **Falsely Classified as Positive**: *I finally caught up to "Starlight" last night on television and all I can say is. . . wow! It's hard to know where to begin -- the incredibly hokey special effects (check out the laser beams shooting out of Willie's eyes!), the atrocious acting, the ponderous dialogue, the mismatched use of stock footage, or the air of earnest pretentiousness that infuses the entire production. This truly is a one-of-a-kind experience,*

and we should all be thankful for that. I nominate Jonathon Kay as the true heir to Ed Wood!

- **Falsely Classified as Negative**: *I'd heard of Eddie Izzard, but had never seen him in action. I knew he was a transvestite, and when I saw he was on HBO one night last summer, I put it on, not knowing how my husband would react. Well, he blew us away. He's better than Robin Williams ever was. He has total control of the audience; when he does the 'Englebert is dead - no he's not', routine, the audience doesn't know what to think by the end*

- **Few shot prompt:**
  - **Truly classified as Positive**: *Not only is this film entertaining, with excellent comedic acting, but also interesting politically. It was made at the end of the Soviet Union, but makes fun of the soviet mentality through and through. The story is set during the early days of the soviet union, and it questions the rationale behind the revolution both in cultural and practical terms...*

    Truly classified as Negative: *I was very unimpressed with Cinderella 2 and Jungle Book 2, but this is possibly worse than both titles. First of all, I didn't like the animation, very Saturday-morning-cartoon, only worse in some scenes...*
  - **Falsely Classified as Negative**: *I've never been a fan of Farrah Fawcett...Until now. She was truly amazing in this movie. The emotion she must have gone through shooting re-take after re-take doesn't bare thinking about. This was a very hard movie to watch, the subject matter is decidedly unpleasant and you feel so helpless just sitting and watching a woman being abused for what seems like an eternity.*
  - Falsely Classified:

- **Instruction-based prompt:**

- **Truly classified as Positive**: *Not only is this film entertaining, with excellent comedic acting, but also interesting politically. It was made at the end of the Soviet Union, but makes fun of the soviet mentality through and through. The story is set during the early days of the soviet union, and it questions the rationale behind the revolution both in cultural and practical terms...*
- **Truly classified as Negative**: *waste of my life, ... the director should be embarrassed. why people feel they need to make worthless movies will never make sense to me. when she died at the end, it made me laugh...*
- **Falsely Classified as negative**: *Bogmeister and others have pretty much nailed this. Shore Leave is really TOS' first attempt at lightweight sci-fi (which they would later perfect with the classic Trouble with Tribbles). It gave both the crew of the Enterprise and its TV viewers a needed respite from the universe threatening consequences of, for example, The Corbomite Manouever...*

שאלות

1. **האם התוצאות שקיבלתם תאמו לציפיות שלכם? הסבירו.**

אחוזי הדיוק שקיבלנו טובות מאוד ועומדות בציפיות שלנו, לאור כך שלא ביצענו Fine-tuning למודל. מה שהיה לא צפוי הוא הקרבה בין תוצאות שלושת הסוגים של הפורמפטים, מכיוון שצפינו לכך שהוספת ההוראה המפורשת למודל, בנוסף לדוגמאות עשויים לשפר את הביצועים שלו, מה שלא קרה בפועל. עם זאת, יש לקחת את התוצאות בערבון מוגבל לאור כך שהמאגר שלנו לא גדול במיוחד ועל כן יכול להיות שלא יהיה מייצג מספיק.

2. **האם התוצאות יהיו שונות אם תתנו יותר מ 2 דוגמאות עבור ה few-shot prompt ? הסבירו.**

לדעתנו, לא יהיה שיפור משמעותי לאחר הוספת יותר דוגמאות. זה נובע מכך שהוספת הדוגמאות ב- Few shot prompt ביחס ל- Zero shot prompt לא הניב תוצאות יותר טובות, מזה אפשר להסיק שהם לא מהוות גורם משמעותי בתהליך.

3. **הסבירו את ההבדל בין Fine-Tuning מסורתי לבין למידה מבוססת פרומפטים (prompt-based learning). במה הם שונים ובמה הם דומים?**

ב- Fine tuning אנחנו מאמנים את המודל מחדש על מאגר הנתונים הספציפי שיש לנו. למשל, לאמן את מודל המשלים מילה חסרה במשפט על סיווג ביקורות של סרטים לקטגוריות של חיובי ושלילי. בלמידה מבוססת פומפטים, לעומת זאת, לא מבוצע אימון נוסף אלא רק עיצוב של הפקודה מחדש, כמו להפוך אותה למפורשת יותר או להוסיף לה דוגמאות, בהסתמך על הידע שכבר נמצא במודל, בניסיון להכוון אותו. הבדל שראוי לציין גם כן הוא ש- Fine tuning דורש יותר כוח חישובי לעומת למידה מבוססת פרומפטים. הדמיון בהם הוא בזה שבשניהם מנסים להשיג תוצאות יותר טובות למשימה כלשהי.

4. **האם תמיד אפשר להחליף fine-tuning ב prompt-based learning?**

לדעתנו לא, מכיוון שהלמידה מבוססת פרומפטים מסתמכת במידה רבה גם על הידע הטמון במודל, ולכן במקרה שהמשימה חדשה בשבילו ולא נחשף אליה בעבר יהיה קשה להכוון אותו לבצע אותה בצורה מיטבית באמצעות פרומפטים ולא יהיה מנוס מלאמן אותו על דאטה חדשה.

5. **כפי שראיתם, ניסוח הפרומפט יכול להשפיע משמעותית על אופי ואיכות התוצאה. דבר זה מקשה מאוד על האיבלואציה של המודלים – עד כמה הם מצליחים במשימה מסויימת, ובהשוואה בין מודלים- בדיקה לאיזה מודל יש ביצועים טובים יותר על אותה משימה. הסבירו מדוע זה מקשה על האיבלואציה והציעו פתרונות אפשריים.**

ההשוואה בין המודלים תהיה קשה כי לא נוכל לדעת אם ההבדל בביצועים הוא כתוצאה מיכולות המודל או הניסוח של הפרומפט, ובשל כך לא נוכל לזהות בקלות בעיות או נקודות חולשה במודל. כמו כן, אנחנו נצפה לכך שמודלים שונים יעבדו באופן שונה עם פרומפטים שונים, מכאן לא נוכל להשתמש בפרומפט אחד לכולם, שיהווה אינדיקטור לטיב המודל.

פתרונות אפשריים הם למשל שימוש בממוצע של כמות גדולה של פרומפטים כמדד, מה שימצע את "הרעש" שתואר למעלה, או שילוב בין Fine-tuning לפרומפטים, או בכלל שימוש בשיטות אחרות שפחות מושפעות מהניסוח כמו F1-Score אם המשימה מתירה כך.

חלק ה : BIAS

1- התמונות שהמודל הפיק לא תאמו לבקשת השתמשת. האם אתם מסכימיעם הקביעה שמדובר ב bias ?

כן מסכימים מדובר בbias , המודל אכן מתקשה לייצר תמונה שבה האישה נוהגת וגבר יושב לצידה למרות שהבקשה ברורה ומדויקת. המודל מתעלם מהבקשה הנדרשת ומייצר תוצאות שמשמרות תבניות מסוימות שאומן עליהם במקרה הזה , הגברים נתפסים כנהגים כשלידם אישה בתדירות יותר גבוהה מאישה נוהגת כשיש גבר לידה, וזה הטיה מגדרית במערכת.

2- מה ככל הנראה הסיבות שגרמו לו להפיק את התוצאות הנ"ל?

1- כנראה שהמודל אומן ע"י מאגר תמונות (מאינטרנט) שבהן לרוב הגבר מופיע כנהג והאישה בתפקיד נוסעת.
2- הלימדה של המודל באופן סטטיסטית, כלומר אם רוב התמונות גברים הם הנהגים, הוא יטה לשמור על אותו דפוס גם אם מבקשים ממנו משהו אחר.
3- בעת ייצרת הבקשה המודל משלים בצורה אוטומטית לפי מה שהוא "מצפה לראות" לפי הנתונים שעליהם אומן ולא מתוך הבנה אמיתית של מושג הנהיגה.

3- הפורמט: "Create an image woman driving and a man is sitting beside her" התוצאה:

Create an image woman driving and a man is sitting beside her

Here is the image of a woman driving with a man sitting beside her. You can view the details in the car interior and their expressions.

אכן נתקלנו באותה בעיה, chatgpt לא הצליח לייצר התמונה הרצויה.

4- שאלת בונוס:

השתמשנו ב chatgpt לייצר אישה שמרימה גבר עם חיבוק.
התחלנו עם הפורמט:
"Create an image of a woman that hug lifts a man"
אחרי כך הפורמט:
"Not like that, Create the image where the woman hug lifts the man"

התוצאות:

Create an image of a woman that hug lifts a man

Here's the image of a woman lifting a man in a hug.

Not like that, Create the image where the woman hug lifts the man

Here's the image of a woman playfully lifting a man off the ground in a hug.

הוא לא הצליח לייצר את הדרוש ואז נסינו עם פורמטים יותר מדויקים:
" You still do it wrong, create an image where a woman is hugging and lifting a man off the ground. The woman has her arms securely wrapped around the man, elevating him."
והפורמט הבא: " You still creating this wrong it should be the other way around, the woman lifting hug the man"

You still do it wrong, create an image where a woman is hugging and lifting a man off the ground. The woman has her arms securely wrapped around the man, elevating him.

Here's the image of a woman hugging and lifting a man off the ground.

You still creating this wrong it should be the other way around, the woman lifting hug the man

Here's the image of a woman lifting a man off the ground in a tight hug.

The woman who lifts hug the man not the opposite

למרות שהפורמטים מדויקים הוא לא מצליח לייצר את הדרוש. אפשר לראות שמתחת לתמונה הוא כתה שייצר אישה מרימה הגבר עם חיבוק למרות שהתמונה שייצר אינה כזאת.

הפורמט הבא:

"The woman who lifts hug the man not the opposite"

The woman who lifts hug the man not the opposite

Here's the image of a woman lifting a man in a hug, with her arms around his back, lifting him off the ground.

הוא עדין לא הצליח לייצר את הדרוש אחרי 5 ניסיונות עם פורמטים שונים.
נסינו פורמטים בעברית: " תיצר תמונה אישה מרימה גבר עם חיבוק "
והפורמט : "האישה מרימה הגבר ולא ההפך"

תיצר תמונה אישה מרימה גבר עם חיבוק

Here's the image of a woman lifting a man off the ground in a warm hug.

האישה מרימה הגבר ולא ההפך

Here's the image of a woman hugging and lifting a man off the ground.

גם כן לא הצליח לייצר את הנדרש.