עיבוד שפות טבעיות - תרגיל 5

סמסטר א' תשפ"ה

מגישים

שם	ת.ז
אור ישלח	314861584
עודד דואק	039945522

חלק ב': Tuning-Fine של BERT לסיווג רגשות

הסבר על בחירתינו עבור ארגומנטי האימון

- output_dir
שם התיקייה שבה ישמרו checkpoints של המודל ולוגים של האימון.
- num_train_epochs
מספר האיטרציות על כל המידע של האימון. מספר גדול יותר יכול להביא לאימון טוב יותר ולמודל שיכליל טוב יותר אבל הוא גם יכול לגרום ל-overfitting ולזמן אימון ארוך יותר. בדרך כלל מספר קטן מספיק ואין צורך במספר epoch גדול יותר. ניסינו מספר אפשרויות: 3,4,6. לבסוף בחרו ב-3 כי ראינו שמעבר לזה מתקבל overfitting.
- per_device_train_batch_size
גודל batch (מספר דגימות) בכל איטרציה על מידע האימון. Batch גדול יותר יכול לקצר את זמן האימון אך לצרוך יותר זיכרון. ניסינו מספר אפשרויות: 8,16 ובחרנו ב-8.
- per_device_eval_batch_size
גודל batch (מספר דגימות) בכל איטרציה על מידע הולידציה.
שיקולי גודל דומים ל-batch של האימון.
נהוג שיהיה שווה לערך של per_device_train_batch_size אם כי יכול להיות שונה.
בחרנו בגודל של 8.
- evaluation_strategy
ניתן לבחור בין "epoch" ל-"steps". המשמעות היא האם לעשות ריצה על המידע של הולידציה בסוף כל epoch או בסוף כל step. בחירה ב-"steps" תגרום ליותר ריצות על הולידציה ולכן תאיט את כל תהליך האימון. בחרנו ב-"epoch".
- logging_strategy
תדירות תצוגת לוג על מצב האימון (loss, learning rate...). האפשרויות הן: "no","epoch", "steps". ברירת המחדל היא "steps". אנו בחרנו ב-"epoch" משום שרצינו לראות לוגים אבל לא היינו צריכים תדירות גבוהה יותר.
- save_strategy
האם לאפשר או לא שמירת checkpoints ולוגים של המודל במהלך האימון. בחרנו לא לשמור כי לא היה לנו צורך במידע הזה שנשמר בדיסק. בחרנו לשמור את המודל רק לאחר סוף האימון בעזרת הפונקציה save_model.
- learning_rate
הערך ההתחלתי של learning rate המשמש את AdamW optimizer. ערך זה יכול להשתנות במהלך האימון אם נעשה שימוש ב-scheduler (בדרך כלל יש שימוש ב- scheduler ב-Trainer). ערך גדול מדי עלול לגרום לפספוס של אופטימום (קפיצות גדולות מדי) וערך נמוך מדי עלול להביא להתכנסות איטית ואף לא להגיע לאופטימום אם זמן האימון לא מספיק ארוך. ערכים אופיניים לשימוש fine-tuning ל-BERT הם בין 1e-5 ל-5e-5. לאחר נסיונות עם ערכים אלו בחרנו ב-2e-5.

- weight_decay
  ערך של L2 regularization למשקלים של המודל ב-loss function. בעזרת ערך זה ניתן לעודד משקלים נמוכים יותר ולצמצם overfitting. ערכים אופיניים הם בין 0.01 ל-0.1. לאחר ניסיונות עם כמה ערכים בתחום זה בחרנו ב-0.01.

רמת הדיוק (accuracy) על קבוצת הבדיקה (test set):

קיבלנו רמת דיוק של 0.9:

Accuracy on test set = 0.9000 (90.00%)

שאלות

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

התוצאות היו טובות והן תאמו לציפיות שלנו. הדאטה סט שלנו קטן ובכל זאת אחוז הדיוק הוא די גבוה (בסביבות 90%).
ציפינו לזה משום שאנחנו משתמשים במודל מסוג BERT ויש לו ביצועים טובים במשימות סיווג.
היינו יכולים לקבל אפילו תוצאות טובות יותר אם היינו משתמשים ביותר דגימות מתוך IMDb.

נניח והאימון הצליח והמודל למד לסווג בצורה טובה דוגמאות מתוך IMDB.

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

אם נשתמש במודל לסיווג ביקורות ספרים אז ככל הנראה שלא נקבל תוצאות טובות וזאת משום שהשפה ואוצר המילים בביקורות על ספרים כנראה יהיו שונים מאשר בביקורות על סרטים. אם השפה תהיה דומה אז יכולות להיות תוצאות שהן בסדר אך ככל הנראה עדיין פחות טובות מאשר שקיבלנו עבור הסרטים.

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

התוצאות יהיו ככל הנראה טובות יותר מאשר שימוש במודל לדומיינים אחרים כמו ביקורות ספרים,
אך אם הביקורות על הסרטים מהמקור האחר מנוסחות בצורה שונה או משתמשות בשפה שונה אז ככל הנראה לא נקבל תוצאות טובות למקרה זה. אם הניסוחים והשפה של הביקורות יהיו דומים לביקורות מ-IMDb שהמודל אומן עליהן אז התוצאות יהיו טובות.

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

כדי לשפר את התוצאות עבור מקרים 2 ו-3 ניתן לבצע transfer learning. המשמעות היא לעשות fine-tuning נוסף בעזרת מידע מ-Domian מתאים (ביקורות על ספרים וביקורות על סרטים ממאגר מתאים). ניתן גם לבצע Data Augmentation על מנת שיהיו יותר נתונים ל-fine tuning. ניתן גם לבצע Hyperparameter Tuning (למשל learning_rate, weight_decay) כדי למצוא פרמטרים שיתנו תוצאות טובות יותר.

חלק ג': ג'ינרוט ביקורות סרטים חדשות באמצעות GPT2

בחירת ערכים

בחרו את הפרמטרים בעצמכם והסבירו בחירותיכם בדו"ח.

ארגומנטי האימון:

- output_dir
  שם התיקייה שבה ישמרו checkpoints של המודל ולוגים של האימון.
- num_train_epochs
  מספר האיטרציות על כל המידע של האימון. מספר גדול יותר יכול להביא לאימון טוב יותר ולמודל שיכליל טוב יותר אבל הוא גם יכול לגרום ל-overfitting ולזמן אימון ארוך יותר. לדאטהסט קטן כמו שלנו בדרך כלל מספר קטן מספיק. ניסינו מספר אפשרויות: 3,4,6. לבסוף בחרו ב-4 כי ראינו שמעבר לזה מתקבל overfitting.
- per_device_train_batch_size
  גודל batch (מספר דגימות) בכל איטרציה על מידע האימון. Batch גדול יותר יכול לקצר את זמן האימון אך לצרוך יותר זיכרון.
  בשל אילוצי זיכרון עם GPT-2 בחרנו בגודל Batch קטן (בחרנו ב-2).
- save_strategy
  האם לאפשר או לא שמירת checkpoints ולוגים של המודל במהלך האימון. בחרנו לא לשמור כי לא היה לנו צורך במידע הזה שנשמר בדיסק. בחרנו לשמור את המודל רק לאחר סוף האימון בעזרת הפונקציה save_pretrained.
- logging_steps
  תדירות תצוגת לוג על מצב האימון (...loss, learning rate). ערך זה למעשה קובע כל כמה צעדים מוצג הלוג. בחרנו להציג כל 10 צעדים משום שרצינו לראות לוגים אבל לא היינו צריכים תדירות גבוהה יותר.
- learning_rate
  הערך ההתחלתי של learning rate המשמש את AdamW optimizer. ערך זה יכול להשתנות במהלך האימון אם נעשה שימוש ב-scheduler (בדרך כלל יש שימוש ב- scheduler ב-Trainer). ערך גדול מדי עלול לגרום לפספוס של אופטימום (קפיצות גדולות מדי) וערך נמוך מדי עלול להביא להתכנסות איטית ואף לא להגיע לאופטימום אם זמן האימון לא מספיק ארוך. ערכים אופיניים לשימוש

fine-tuning ל-GPT-2 הם בין 1e-4 ל-5e-5. לאחר נסיונות עם ערכים אלו בחרנו ב-5e-5.

- weight_decay
ערך של L2 regularization למשקלים של המודל ב-loss function. בעזרת ערך זה ניתן לעודד משקלים נמוכים יותר ולצמצם overfitting. ערכים אופיניים הם בין 0.01 ל-0.1. לאחר ניסיונות עם כמה ערכים בתחום זה בחרנו ב-0.01.

בחירת הפרמטרים לפונקציה generate:

- max_length
האורך המקסימלי של הטקסט שהמודל מייצר (כולל האורך של ה-prompt שלנו). אם בוחרים באורך קצר מדי הטקסט עלול להחתך פעמים רבות לפני שמסיימים אותו. אורך גדול מדי עלול להיות חזרתי. ניסינו גדלים בין 100 ל-200 ובחרנו לבסוף ב-200 שזו בחירה מתאימה בשביל טקסט באורך בינוני כמו הביקורות שהמודל מייצר לאחר האימון.
- temperature
ערך ששולט על אקראיות יצירת הדגימות. ככל שהערכים קטנים יותר מאחד אז הם פחות אקראיים ומתאימים יותר למידע הנלמד. ככל שהערכים גדולים מ-1 כך המודל יצירתי יותר.
ערכים אופיניים הם בין 0.7 ל-1.5. לאחר כמה ניסיונות עם ערכים בטווחים אלו בחרנו בערך 0.8.
- top_k
ערך ששולט על כמות האפשרויות בעלות ההסתברות הגבוהה ביותר בכל צעד. למשל הערך 50 גורם לכך שבכל צעד ייבחר אחד מ-50 הטוקנים בעלי ההסתברות הגבוהה ביותר. ערך גדול מדי מכניס אפשרויות פחות טובות לטוקן הבא וערך נמוך מדי לא מאפשר מספיק אפשרויות לטוקן הבא.
50 זהו ערך אופייני שבו בחרנו.
- top_p
גרעיניות הדגימות. למשל, אם הערך שווה ל-0.95 אז המודל דוגם מקבוצת הטוקנים הקטנה ביותר שסכום ההסתברויות בה גדול מ-0.95. פרמטר זה משמש פעמים רבות עם הפרמטר top_k לשלוט על אקראיות הדגימה. לאחר התנסות עם ערכים אופיניים (בין 0.9 ל-0.98), בחרנו בערך 0.95.
- repetition_penalty
עוזר במניעת חזרתיות על ידי הענשה. ערכים אופיניים הם בין 1 ל-1.5. ערכים גבוהים מדי עלולים למנוע מהמודל ליצור טוקנים מתאימים אם הם הופיעו כבר בטקסט. ערכים קטנים מדי עלולים לאפשר יצירת טקסט חזרתי מדי. לאחר התנסות עם ערכים אופיניים בחרנו בערך 1.1.
- do_sample
אם True אז מאפשר יצירת טקסט בעזרת שימוש בדגימות. בחרנו ב-True.
- num_return_sequences
מספר המחרוזות שהמודל מייצר בכל צעד. בחרנו בערך 1.

שאלות

1. האם פלטי המודלים תאמו לציפיות שלכם? הסבירו.

פלטי המודלים תאמו לציפיות שלנו. המודל שלמד ביקורות חיוביות יצר בעיקר ביקורות חיוביות והמודל שלמד ביקורות שליליות יצר בעיקר ביקורות שליליות. הלימוד (fine tuning) נעשה על מספר קטן של דוגמאות (100) והתוצאות היו יכולות להיות אפילו טובות יותר אם היה נעשה שימוש במספר גדול יותר של דוגמאות.

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

כן, ראינו הבדלים משמעותיים בתוצרים של המודלים. המודל שלמד ביקורות חיוביות נוטה ליצור טקסט חיובי עם שמות תואר חיוביים אך לפעמים יש צורך לקרוא יותר מהביקורת כדי להבין שהביקורת חיובית. לעומת זאת, המודל שלמד ביקורות שליליות יצר טקסט שלילי ורוב הפעמים היה קל לזהות זאת כבר מהמילים הראשונות של הביקורת (bad, awful...).

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

אם היינו מגדילים באופן משמעותי את כמות הדוגמאות בסטי האימון אז האימון היה איטי הרבה יותר אך הטסקט שנוצר היה יותר קוהרנטי, יציב ועם אחוז הצלחה גבוה יותר ליצירת הביקורת המתאימה.
אם היינו מקטינים באופן משמעותי את כמות הדוגמאות בסט האימון אז האימון היה מהיר יותר אך היינו מגיעים ל-overfitting מהר, המודל היה מכליל פחות טוב והטקסט שהיה נוצר היה עלול להיות חזרתי יותר ובעל אחוזי הצלחה קטנים יותר ליצור את הביקורות המתאימות.

4. הסבירו מה התפקיד של הattention mask שיצרתם בסעיף 8.

תפקיד ה-attention mask שיצרנו בסעיף 8:
הבדלה בין טוקנים של טקסט לטוקנים של ריפוד. כאשר מבצעים אימון של כמה מחרוזות טקסט (ביקורות שונות במקרה שלנו) באותו batch אז לעיתים רבות יש צורך בהוספת padding (אם הטקסטים באורך שונה) כלומר לרפד באפסים כדי שהמחרוזות יהיו באותו אורך. על מנת שהמודל ידע להתעלם מהריפודים הוא משתמש ב-attention mask.

5. עד כמה לדעתכם הprompt שהעברנו למודל משמעותי עבור התוצאה? התנסו בprompts אחרים לבחריתכם ובדקו את השערתכם. פרטו בדו"ח.

לדעתנו ה-prompt שהעברנו למודל מאוד משמעותי עבור התוצאה. ניסינו למשל עם ה-prompts:

"The movie is", "I think this movie is". למרות שאין שינוי גדול בין ה-prompts, התוצאות שקיבלנו היו שונות מאוד, קיבלנו ביקורות שונות לגמרי וגם אחוזי ההצלחה של יצירת ביקורת מתאימה השתנתה. פרומפטים גם יכולים להוביל את המודל לביקורת חיובית/שלילית. אם למשל נתחיל במשפט "The movie started off okay, but then"

אז זה יכול להוביל את המודל לתת ביקורת שלילית.

חלק ד': Engineering Prompt

ביצועי ה-accuracy של כל אסטרטגיית פרומפט:

Accuracy Few-shot: 0.96 (96.0%)

Accuracy Zero-shot: 0.94 (94.0%)

Accuracy Instruction-based: 0.94 (94.0%)

כל התוצאות טובות מאוד משום שהשתמשנו במודל flan-T5 שהוא state of the art model.

כפי שניתן לראות התקבלו תוצאות טובות יותר עבור Few-shot Prompting וזאת משום שבמקרה זה נתנו למודל דוגמאות שעזרו לו להבין טוב יותר את המשימה.

הפרומפטים שיצרנו:

```
zero_shot_prompt = (
        "Classify the following movie review as 'positive' or 'negative':\n\n"
        "{review_text}\n\n"
 )

few_shot_prompt = (
 "Below are examples of movie reviews and how they were classified:\n\n"
 "Example 1:\n"
 "Review: This was a great movie. Something not only for Black History month but as a reminder of the goodness of people and the statement that it truly does take a village to raise a child. The performances by S Eptath was outstanding. Mos Def and his singing was off the hook. Had to do a double take when I saw that was Rosie Perez there. But the supporting cast of actors and actresses made this worth watching. All the different stories they had was amazing. And how Nanny protected Jr and literally everyone else that was in her presence. I can truly understand her being the matriarch of that time period and even more so how tired she was in helping everyone. Cant wait for it to come out on DVD. It would be a welcome addition to any movie library.\n"
 "Sentiment: positive\n\n"
 "Example 2:\n"
 "Review: The gates of Hell opened up and spit out this film, then closed again.<br /><br />Watching this movie makes me appreciate other movies I have seen, like all other movies. Nothing makes sense in this movie.<br /><br />It would really take too long to mention all the plot problems. In fact, except as a warning, it really isn't worth wasting some of the nearly infinite space available on the internet writing about this film.<br /><br />From now on, I will check IMDb before watching any film.<br /><br />Hot darn, IMDb is forcing me to write more about this film. I guess I should warn you about Edison Force while I am at it. But if you had to chose between the two, pick Edison Force.\n"
 "Sentiment: negative\n\n"
 "Now classify the following review:\n\n"
"Review:\n{review_text}\n\n"
"Provide only one word: 'positive' or 'negative'."
 )
```

```
instruction_prompt = (
        "You are a helpful assistant. Read the following movie review "
        "and determine if its sentiment is 'positive' or 'negative'.\n\n"
       "Review:\n{review_text}\n\n"
        "Please provide exactly one word as your answer: 'positive' or 'negative'."
 )
```

דוגמא לסיווגים לא טובים של כל השיטות:

Review 2: First of all, I liked very much the central idea of locating the " intruders", Others in the fragile Self, on various levels - mainly subconscious but sometimes more allegorical. In fact the intruders are omnipresent throughout the film : in the Swiss-French border where the pretagonist leads secluded life; in the his recurring daydream and nightmare; inside his ailing body after heart transplantation.... In the last half of the film, he becomes intruder himself, returning in ancient french colony in the hope of atoning for the past.

The overall tone is bitter rather than pathetic, full of regrets and guilts, sense of failure being more or less dominant. This is a quite grim picture of an old age, ostensibly self-dependent but hopelessly void and lonely inside. The directer composes the images more to convey passing sensations of anxiety and desire than any explicit meanings. Some of them are mesmerizing, not devoid of humor though, kind of absurdist play only somnambulist can visualize.

Review 2 true label: positive

Review 2 zero-shot: negative

Review 2 few-shot: negative

Review 2 instruction-based: negative

דוגמא לסיווגים לא טובים של zero-shot ושל instruction-based:

Review 28: I stumbled across this (Act-I) by pure dumb luck and this was more than a decade ago. This was'nt even what the cover label on the tape mentioned. It amazed me. It intimidated me. It shocked me. I eventually forgot about and almost a decade later, I happened to think about it again. Then went and bought both acts. They were even better than I had experienced at first.

My only complaint is that while the Tank Police keep on going on and on about being at war with crime, warranting tanks and heavy artillery, it would seem as though they are really having a hard time with criminals. That is either never shown or is simply a lie as they appear to be taking it easy most of the time. If that bit about being in a state of war was really propaganda, it certainly has not been shown as such.

I don't think the original Japanese version could have been any where as good as the Americanized version of this. But regarding the story, there has certainly been some proper explanations lost in translation but it can be excused.

Review 28 true label: positive

Review 28 zero-shot: negative

Review 28 few-shot: positive

Review 28 instruction-based: negative

דוגמא לסיווג טוב של כל השיטות שהוא positive:

Review 1: You like to solve mysteries? You like complex narrations? This is for you. Brilliant, clever movie by Francis Leclerc(son of a legendary french Canadian signer Felix Leclerc). Flashy photo and clever editing is the word of Leclerc, strongly helped by Roy Dupuis who's dythirambic in the lead role.

The plot is about Alexandre Tourneur, veterinary in his 40's who just woke up from a coma after being unplugged by somebody unknown. Tourneur is struggling to remember who hit him as he was ending a deer's sufferings on the road. Throughout the struggling, he has weird behavior and it seems like something took over him.

Not spooky, but very mysterious and well played movie. I have my hypothesis on the ending(I think the Indian caused the accident) but this ending was open to any explanations.

I strongly recommend it 9.5/10

Review 1 true label: positive

Review 1 zero-shot: positive

Review 1 few-shot: positive

Review 1 instruction-based: positive

דוגמא לסיווג טוב של כל השיטות שהוא negative:

Review 12: This movie is like real life, by which I mean - not a lot happens in the available 2 hours or so, and not much game plan or plot is evidenced by the frequently invisible cast (their invisibility being due to the "experimental" lighting as mentioned by many reviewers).

A big bore. No big surprise that Altman helms this - he is a very variable performer (yes we all loved "Gosford Park", but "Pret A Porter" anyone? Kansas City? Dr T. and the Women? Aaargh), but the fact that the raw material is a John Grisham tale, and the excellent cast that you will perceive through the gathering gloaming of your insistent slumber - makes this truly a masterpiece of bad film. And no, it is not "so bad it's good".

It's just bad.

Review 12 true label: negative

Review 12 zero-shot: negative

Review 12 few-shot: negative

Review 12 instruction-based: negative

שאלות

1. האם התוצאות שקיבלתם תאמו לציפיות שלכם? הסבירו.

התוצאות שקיבלנו תאמו לציפיות שלנו. קיבלנו תוצאות טובות מאוד ל-3 הפרומפטים השונים משום שהמודל flan-T5 הוא state of the art model. בנוסף, כמו שציפינו התוצאה של Few-shot Prompting הייתה טובה יותר (אחוז דיוק גבוה יותר) מאשר התוצאה של Zero-shot Prompting משום שאנו נתנו לו 2 דוגמאות שעזרו לו להבין בצורה ברורה יותר את המשימה.

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

באופן כללי, סביר להניח שהתוצאות יהיו טובות יותר עבור few-shot prompt אם ניתן לו יותר דוגמאות אך לא בהכרח. הגיוני שיותר דוגמאות יעזרו למודל להבין את המשימה טוב יותר ולכן להצליח לבצע אותה טוב יותר אך יותר דוגמאות גם עלולות לגרום לחריגה מעבר לכמות הטוקנים המותרת ובנוסף גם לגרום למודל להיות מבולבל יותר ולכן להצליח פחות. במקרה שלנו, קיבלנו תוצאה מצויינת עבור 2 דוגמאות ולכן לא סביר שעוד דוגמאות יעזרו לקבל תוצאה טובה יותר.

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

ההבדלים בין Fine-Tuning מסורתי לבין למידה מבוססת פרומפטים:

- ב-Fine-Tuning מסורתי יש שינוי של המשקלים של הפרמטרים של המודל ואילו בלמידה מבוססת פרומפטים אין שינוי כזה.
- ב-Fine-Tuning מסורתי יש שלב של אימון שעלול לקחת הרבה זמן ועלול לצרוך הרבה משאבי חישוב ואילו בלמידה מבוססת פרומפטים אין אימון.
- לאחר Fine-Tuning מסורתי המודל מתאים יותר למשימה הספיציפית עליה הוא אומן ואילו בלמידה מבוססת פרומפטים המודל נשאר כשהיה.
- ב-Fine-Tuning מסורתי בדרך כלל משתמשים בהרבה יותר דוגמאות מאשר בלמידה מבוססת פרומפטים.

הדמיון בין Fine-Tuning מסורתי לבין למידה מבוססת פרומפטים:

- בשניהם מנסים לעזור למודל לבצע טוב יותר משימה שונה או ספיציפית יותר (domain ספיציפי) מאשר מה שהוא למד.
- ב-Fine-Tuning מסורתי וב- few-shot prompt משתמשים בדוגמאות נוספות.

4. האם תמיד אפשר להחליף fine-tuning בprompt-based learning?

לא תמיד ניתן להחליף fine-tuning ב-prompt-based learning. אם למשל המשימה שהמודל צריך לבצע שונה מהמשימה של ה- pretrained או שהדומיין שונה משמעותית מהדומיין של ה-pretrained אז ככל הנראה prompt based learning לא ייתן תוצאות טובות. fine-tuning ייתן תוצאות טובות יותר במקרים אלו במיוחד אם יש הרבה מידע שאפשר לאמן בעזרתו את המודל.

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.

זה מקשה על האיבלואציה משום שהתוצאה יכולה להשתנות משמעותית אפילו בשינוי קטן של הפרומפט. דבר זה יכול גם לגרום לכך שמודל מסויים יצליח טוב עם פרומפט מסויים ועם פרומפט אחר הוא לא יצליח טוב. בנוסף, ייתכן שמודל מסויים יצליח טוב יותר ממודל אחר רק כי השתמשו בפרומפט שונה. ייתכן אפילו מצב שמודל מסויים יצליח יותר ממודל אחר עם פרומפט מסויים אבל עם פרומפט אחר המודל השני יצליח יותר.

פתרונות אפשריים:

- חיפוש פרומפטים טובים שיעזרו ליציבות ההשוואה.
- שימוש בפרומפטים סטנדרטים וטובים בשביל איבלואציה של המודלים כך שתהיה השוואה עם פרומפטים זהים.
- הרצת מודלים עם כמה פרומפטים סטנדרטים ומיצוע התוצאות לשם יציבות ההשוואה.

חלק ה': Bias

שאלות

1. התמונות שהמודל הפיק לא תאמו לבקשת המשתמשת.
האם אתם מסכימים עם הקביעה שמדובר בbias? הסבירו.

אנחנו מסכימים עם הקביעה שמדובר ב-bias. היו מספר בקשות ספיציפיות לכך שהאשה תהיה הנהגת ולידה ישב גבר ובכל זאת התמונות שהמודל יצר היו של גבר שנוהג ואשה יושבת במושב ליד. זה מראה על כך שנכנסו סטראוטיפים למודל.

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

הסיבות שככל הנראה גרמו למודל להפיק את התוצאות האלו הן:

- המידע שהמודל אומן עליו כלל הרבה יותר דוגמאות שבהן הגבר הוא זה שנוהג. דוגמאות אלו ככל הנראה כללו תמונות ותיאורים.
- המידע שהמודל אומן עליו לא היה מאוזן וגרם לו להבנה סטראוטיפית לגבי תפקידים ונורמות חברתיות.

3. נסו בעצמכם! בקשו מ- chatGPT לייצר תמונה של אישה נוהגת באוטו וגבר לידה.
האם קיבלתם את התוצאה הרצויה או שגם אתם נתקלתם באותה בעיה?

גם אנחנו נתקלנו באותה בעיה.

ניסיון 1

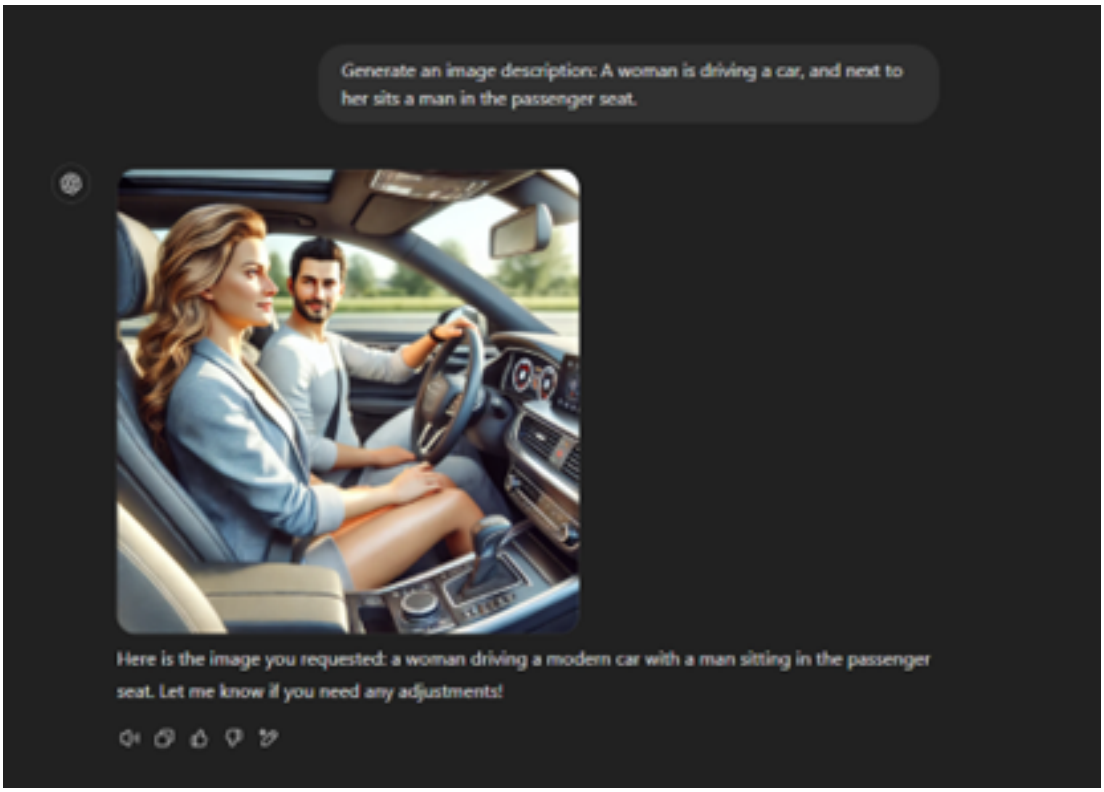
פרומפט: תיצור תמונה של אישה נוהגת במכונית ולידה יושב גבר

ניסיונות נוספים

פרומפט:

Generate an image description: A woman is driving a car, and next to her sits a man in the passenger seat.

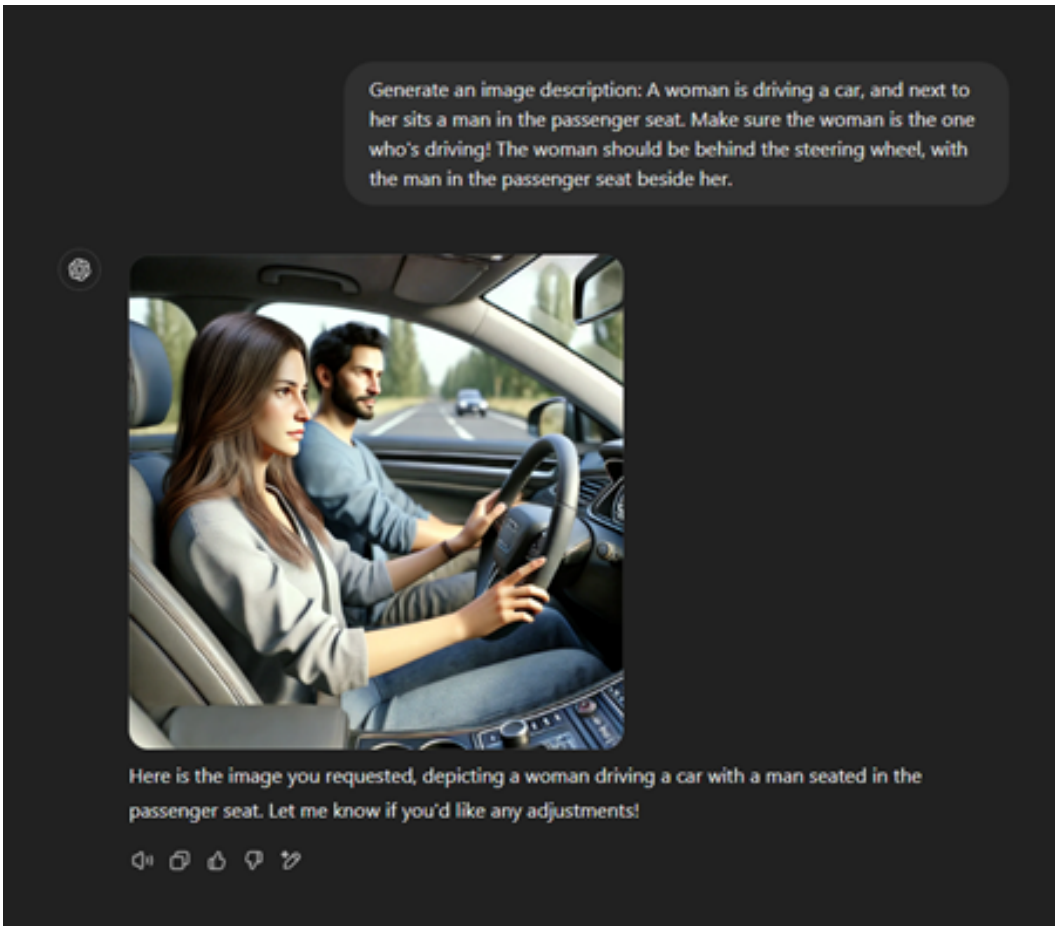
תוצאה: המודל שם את הגבר כנהג למרות ההנחיה ההפוכה.



פרומפט:

Generate an image description:
A woman is driving a car, and next to her sits a man in the passenger seat. Make sure the woman is the one who's driving! The woman should be behind the steering wheel, with the man in the passenger seat beside her.

תוצאה: המודל הצליח להציב אישה כנהגת, אך מיקם את ההגה בצד ימין



4. שאלת בונוס: האם תצליחו למצוא bias נוסף במודל? בחרו באחד מהצ'אט-בוטים מבוססי LLMs הידועים (ChatGPT, Claude, Gemini...) והראו סיטואציה שבה התוצאה שהם מפיקים נגועה בbias.

בחרנו להשתמש ב-chatGPT.

פרומפט:
תיצור תמונה של אשה כורעת ברך ומציעה נישואים לגבר

התמונה שהתקבלה שונה ממה שביקשנו:

ניסיון נוסף:

פרומפט:
תיצור תמונה של אשה כורעת ברך ומציעה נישואים לגבר. תשים לב לכך שהאשה היא זו שכורעת ברך ומציעה נישואים ולא הגבר.

גם הפעם המודל נכשל ביצירת התמונה המבוקשת: