https://github.com/hoshenlugasi/Anlp_ex1

- ANLP EX1

שאלה 1:

1. **Word in Context – WIC**

הדאטאסט בודק האם מילה משותפת שמופיעה בשני משפטים שונים משמשת באותה משמעות או לא. זו תכונה פנימית של הבנת שפה, כי היא בודקת האם המודל מבין הבדלים סמנטיים עדינים בין שימושים במילים. כלומר את היכולת הפנימית של הבנת משמעות של מילים בהקשר מסוים.

2. **Commitment Bank – CB**

הדאטאסט בודק האם טקסט מסוים נובע, סותר או ניטרלי ביחס לטקסט אחר . מודד את היכולת הפנימית של זיהוי יחסי הסקה טקסטואלית ובוחן באופן ישיר את יכולת ההבנה הלוגית-סמנטית של יחסים בין משפטים. זו היא תכונה של הבנת שפה ללא תלות במשימות חיצוניות.

3. **WSC (Winograd Schema Challenge)**

זהו דאטאסט שמודד את היכולת הפנימית של המודל לבצע coreference resolution . כלומר, לקבוע למה מתייחסים כינויי גוף בהקשרים דו-משמעיים שמצריכים הבנה של ההקשר, התחביר והסמנטיקה. היכולת לזהות את הרפרנט הנכון לא נשענת על ידע חיצוני, אלא על תכונה בסיסית של הבנת שפה, ולכן זהו מדד לתכונה פנימית של הבנת שפה.

שאלה 2:

אתאר מתודות שונות שהמשותף לכולן הוא שיפור דיוק המודל באמצעות הגדלת אורך הפלט של המודל או הגדלת כמות התשובות שהמודל מייצר עבור אותה תשובה (ולא באמצעות הגדלת כמות הדאטא שעליו המודל מתאמן או שינוי גודל המודל).

Self-Consistency

- **תיאור** : יצירת מספר תשובות שונות בעזרת הרצה מספר פעמים ודגימת כמה תשובות, ואז בחירת התשובה הנפוצה ביותר ("majority voting").
- **יתרונות** :מפחית שגיאות רנדומליות, משפר דיוק בבעיות מורכבות, מצריך שינוי רק בזמן הסקה ללא צורך באימון מחדש.

- **צווארי בקבוק חישוביים** :דורש הרצות מרובות של המודל ולכן זמן ריצה וחישוב גבוה יותר, שימוש גבוה בזכרון בגלל אחסון מקבילי של התשובות.
- **מקבול** :ניתן להקביל בקלות את יצירת התשובות השונות, שכן הן בלתי תלויות זו בזו.

Verifiers

- **תיאור** : שימוש בכלים שבוחנים את איכות ונכונות הפלט מהמודל המקורי. כלומר, להגיד האם תשובה של מודל היא נכונה או לא. לדוגמא בתכנות, כלים אוטומטים כמו unit tests.
- **יתרונות** : מאפשר סינון של תשובות שגויות, או בחירת התשובה הטובה ביותר מבין כמה אפשרויות, יעיל במיוחד בתחומים עם יכולת אימות (כמו תכנות), ניתן לשלב עם גישות אחרות.
- **צווארי בקבוק חישוביים** :הוספת שלב חישובי נוסף לתהליך, עלויות חישוביות של המוודא עצמו, מורכבות בבניית מוודאים לתחומים מסוימים. צריכת זכרון מוגברת – צריך להחזיק לפעמים כמה פתרונות עד לסיום שלב האימות.
- **מקבול** :ניתן להקביל את יצירת ובדיקת התשובות השונות, אך לא תמיד את תהליך האימות עצמו.

Chain of Thought (CoT)

- **תיאור** :שיטה שבה המודל מייצר סדרת צעדי הנמקה מפורשים לפני שמגיע לתשובה הסופית, במקום לענות ישירות. במקום להגדיל את כמות הדאטא שעליו מודל אומן, להגדיל את אורך הפלט שלו ולאפשר שלו לו לתכנן את התשובה שלו.
- **יתרונות** : המודל חושב בצעדים הדרגתיים, מאפשר לו לפרק בעיות מורכבות בצורה טובה יותר. כך, משפר ביצועים במשימות המצריכות הסקה מורכבת. מאפשר מעקב אחר תהליך החשיבה ואיתור שגיאות.
- **צווארי בקבוק חישוביים** : צריכת טוקנים רבים יותר (הפלט ארוך משמעותית), עלייה בזמן החישוב, שימוש מוגבר בזיכרון עבור שמירת הקשר ארוך יותר.
- **מקבול** : לא ניתן להקביל את התהליך, משום שנוצרת שרשרת אחת סדרתית שבה יש תלות בין הטוקנים.

LEAST TO MOST

- **תיאור** : שיטה איטרטיבית לפתרון בעיות מורכבות: במקום שהמודל יפתור הכל בבת אחת, הוא מקבל את הבעיות שלב שלב, מזהה את תתי הבעיות, פותר אותן אחת אחת, ורק אז פותר את השאלה הראשית.
- **יתרונות :**מאפשר למודל לפרק בעיה מורכבת ולפתור כל שלב בפשטות יחסית. משפר ביצועים בבעיות reasoning מורכבות.
- **צווארי בקבוק חישוביים**: כל שלב דורש קריאה נפרדת למודל, עולה בזמן ובהוצאות חישוביות.
- **מקבול** : לא, התהליך תלוי בשלבים הקודמים (איטרטיבי וסדרתי).

SELF-ASK

- **תיאור :** שיטה שבה המודל שואל את עצמו שאלות משנה שיכולות לעזור לו לפתור את השאלה הראשית. הוא מבצע self-querying, עונה לעצמו, ואז ממשיך (מכווינים לזה באמצעות פרופמט מתאים).
- **יתרונות :** עוזר למודל "לחשוב בקול רם" ולבנות פתרון מורכב ממספר צעדים קלים יחסית.
- **צווארי בקבוק חישוביים:** אורך פלט ארוך, חישוב יקר בהפעלה אחת.
- **מקבול :** לא נדרש, הכל קורה באותה ריצה.

Increasing Output Length

- **תיאור** :מתן יותר "מרחב טוקנים" למודל לפתח את חשיבתו, עם פחות הגבלה על אורך הפלט. מאפשר למודל לתכנן את התשובה שלו, לעשות backtracking ולתקן אותה במהלך כתיבת הפלט, ולעשות self-evaluation לעצמו בזמן כתיבת הפלט.
- **יתרונות** :מאפשר למודל לפתח תהליכי חשיבה עמוקים יותר, לתכנן את התשובה שלו, לתקן את עצמו בזמן ריצה. משפר יכולת להתמודד עם בעיות מורכבות, קל ליישום.
- **צווארי בקבוק חישוביים** :עלייה בזמן חישוב ביחס לאורך הפלט, שימוש מוגבר בזיכרון ביחס לאורך הפלט, עלול לגרום ל"הליכה מסביב" ללא התקדמות.
- **מקבול** :לא ניתן להקביל את יצירת הטוקנים עצמם.

Tree of Thought

- **תיאור** :המודל מפתח מספר אפשרויות במקביל בכל שלב החלטה, ויוצר מבנה עץ של תהליכי חשיבה. כל ענף הוא שלב בתהליך החשיבה. התשובה הסופית היא העלה עם הציון הגבוה.
- **יתרונות** : חיפוש רחב שבו יש אפשרות לאתר את הפתרון הטוב ביותר. יכולת לחקור מסלולי פתרון חלופיים, שיפור בבעיות מורכבות שדורשות תכנון לעומק, אפשרות לחזור לאחור כשזוהה מסלול לא מוצלח.
- **צווארי בקבוק חישוביים** : צריכת משאבים גבוהה יותר בצורה משמעותית בהתאם לעומק העץ הן של זמן ריצה והן של זכרון אחסון, מורכבות באסטרטגיית חיפוש.
- **מקבול** :ניתן להקביל את חקירת הענפים השונים של העץ.

Increased Compute Budget

- **תיאור** :שימוש במודל קטן כדי לייצר הרבה פלטים, ואז לבחור את הכי טוב באמצעות מוודא טוב, במקום להריץ פעם אחת עם מודל גדול.
- **יתרונות** : משפר דיוק במקרה של verifier טוב, לעומת עלות זהה בהרצה יחידה של מודל גדול.
- **צווארי בקבוק חישוביים** : דורש יותר משאבי אחסון של מספר תשובות עד לשלב הווידוא, דורש הרבה הרצות של המודל (זמן ריצה גדול), תלוי מוודא איכותי.
- **מקבול** : כן, ניתן להריץ כל דוגמה ואת האימות שלה באופן מקבילי.

סעיף ב

1. בהינתן שיש לי GPU יחיד עם הרבה זיכרון, אך ללא יכולת להריץ את המודל במקביל מספר פעמים, הייתי בוחרת בשילוב של **Chain of Thought reasoning** עם **הגדלת אורך הפלט (Increasing Output Length)** לקבלת פלט שמשתמש **ב-Planning** , **Backtracking, Self-Evaluation.**
שיטה זו מאפשרת למודל להפעיל תהליך חשיבה מדורג, שבו הוא מנסח סדרת טיעונים והסקות עד להגעה לתשובה הסופית, והכול בתוך פלט יחיד. במקום להפעיל את המודל שוב ושוב, כמו בself-consistency , אני מנצלת את הזיכרון הגבוה של ה־GPU כדי לאפשר לו להפיק פלט ארוך ומעמיק ,שכולל תכנון של הפתרון, תיקון עצמי (backtracking) , ואף הערכה עצמית של השלבים לאורך הדרך(self) evaluation).

גישה זו הוכחה כאפקטיבית במשימות שדורשות reasoning מורכב, ומספקת איזון בין ביצועים לבין מגבלות החומרה הקיימות, שכן היא משפרת את דיוק הפלט בלי צורך במקבול.

<u>חלק פרקטי:</u>

2. כן, הקונפיגורציה שהשיגה את הדיוק הכי גבוה על סט הולידציה השיגה גם את הדיוק הכי גבוה בסט המבחן (1 הכי נמוך, 3 הכי גבוה).
   1. epoch_num: 3, lr: 0.01, batch_size: 32, eval accuracy: 0.6838, test accuracy: 0.66493
   2. epoch_num: 1, lr: 0.00001, batch_size: 16, eval accuracy: 0.7181, test accuracy: 0.72522
   3. epoch_num: 2, lr: 0.00001, batch_size: 16, eval accuracy: 0.7892, test accuracy: 0.78087

3. לאחר ניתוח הדוגמאות בהן המודל הטוב הצליח והמודל הרע נכשל, זיהיתי מספר דפוסים חוזרים בדוגמאות שהיו קשות יותר למודל החלש:
   1. **עמידות להבדלים מספריים קטנים או זניחים**

המודל הטוב ידע להתעלם מהבדלים כמותיים קטנים כאשר שאר המידע סמנטית זהה, והבין שהמשפטים מתייחסים לאותו אירוע או תיאור, לעומת המודל הראשון שהתקשה וטעה.

דוגמאות:

1 – המשפטים מציינים כמויות שונות (700 לעומת 600 חיילים), אך מתארים את אותו חוסר מוכנות של כוחות האו"ם.

```
[71]
Sentence 1: While there were about 700 Uruguayan UN peacekeepers in Bunia , they were " neither trained nor equipped " to deal with inter-ethnic violence , Mr Eckhard said .
Sentence 2: The UN has approximately 600 troops in the town , but they were " neither trained nor equipped " to deal with inter-ethnic violence , Mr Eckhard said .
True label: 1, Best pred: 1, Bad pred: 0
```

2 – משפט אחד מזכיר 1,750 איש שהתפנו לפני 9:00, והשני מדבר על 1,752 איש שהתפנו ב־8:45 – אך בשני המקרים הכותב מדווח על אותו פינוי, רק עם פירוט שונה.

```
Sentence 1: At least 1,750 people were ordered to evacuate their homes shortly before 9 a.m. as a precaution , said Steve Powers , Marquette County administrator .
Sentence 2: At least 1,752 people were ordered to evacuate their homes in the north part of Marquette about 8 : 45 a.m. EDT , said Steve Powers , Marquette County admi
True label: 1, Best pred: 1, Bad pred: 0
```

2. **הבנת זהות למרות הבדל ברמת פירוט או הכללה**

המודל הטוב ידע לזהות שהמידע זהה גם כאשר אחד המשפטים כלל יותר פרטים מזה השני, לעומת המודל הפחות טוב שהתקשה וטעה.

דוגמאות:

1 – אחד המשפטים כולל רק "השופט", והשני מפרט את שמו ותפקידו, אבל המשמעות לא משתנה.

```
Sentence 1: " There 's no reason for you to keep your skills up , " the judge told the convicted crack cocaine kingpin .
Sentence 2: " There 's no reason for you to keep your skills up , " U.S. District Judge J. Frederick Motz told McGriff after he was sentenced .
True label: 1, Best pred: 1, Bad pred: 0
```

2 כאן המשפט הראשון כללי יותר (who said she wouldn 't comment on personnel issues) בעוד שהשני מתאר פעולה קונקרטית יותר (who didn 't immediately return a call) המודל הטוב הבין שמדובר באותו אירוע או תפקיד בעוד שהמודל הפשוט לא הבחין בכך.

```
Sentence 1: The main psychologist , Dwight Close , referred questions to an agency spokeswoman , who said she wouldn 't comment on personnel issues .
Sentence 2: The main psychologist in the Rodriguez case , Dwight Close , referred questions to an agency spokeswoman , who didn 't immediately return a call .
True label: 1, Best pred: 1, Bad pred: 0
```