NLP and Transformer Models - 3 מטלה

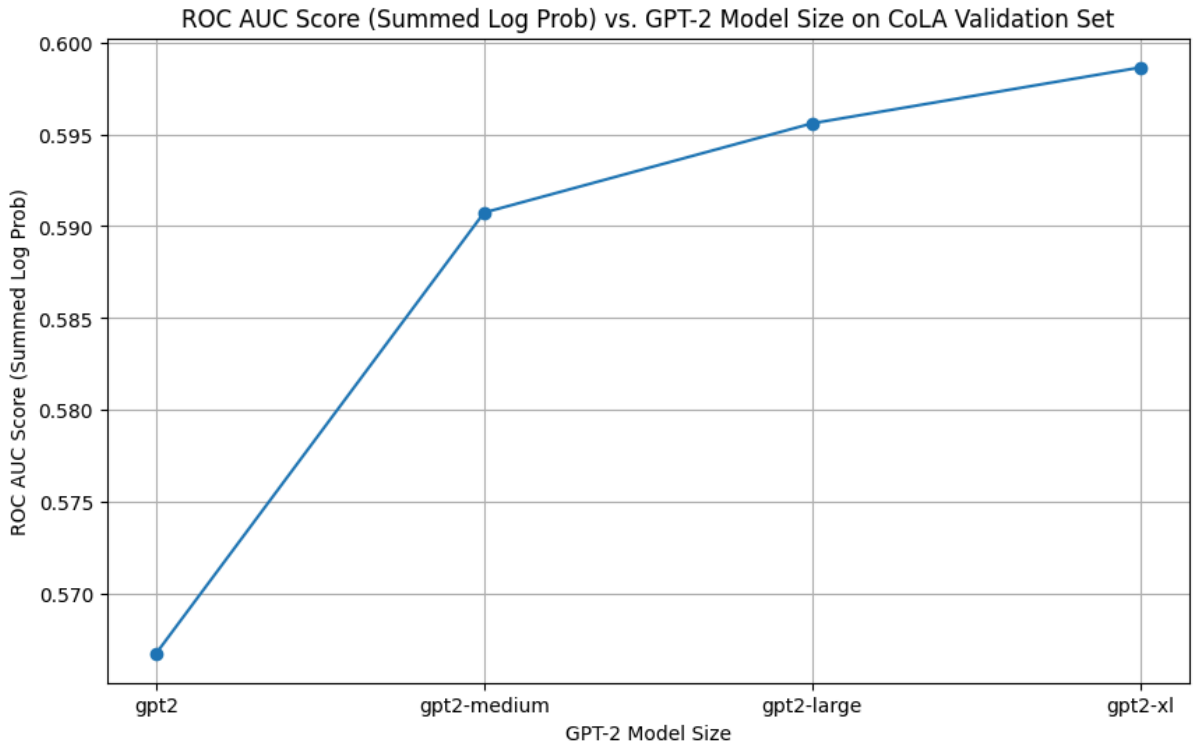
Part 1: Evaluating Autoregressive Language Models for Grammatical Acceptability

בחלק זה של המטלה בחנו את יכולתם של מודלים אוטו-רגרסיביים של שפה לחזות שיפוטים אנושיים של תקינות דקדוקית באנגלית, תוך שימוש במערך הנתונים CoLA (Corpus of Linguistic Acceptability).

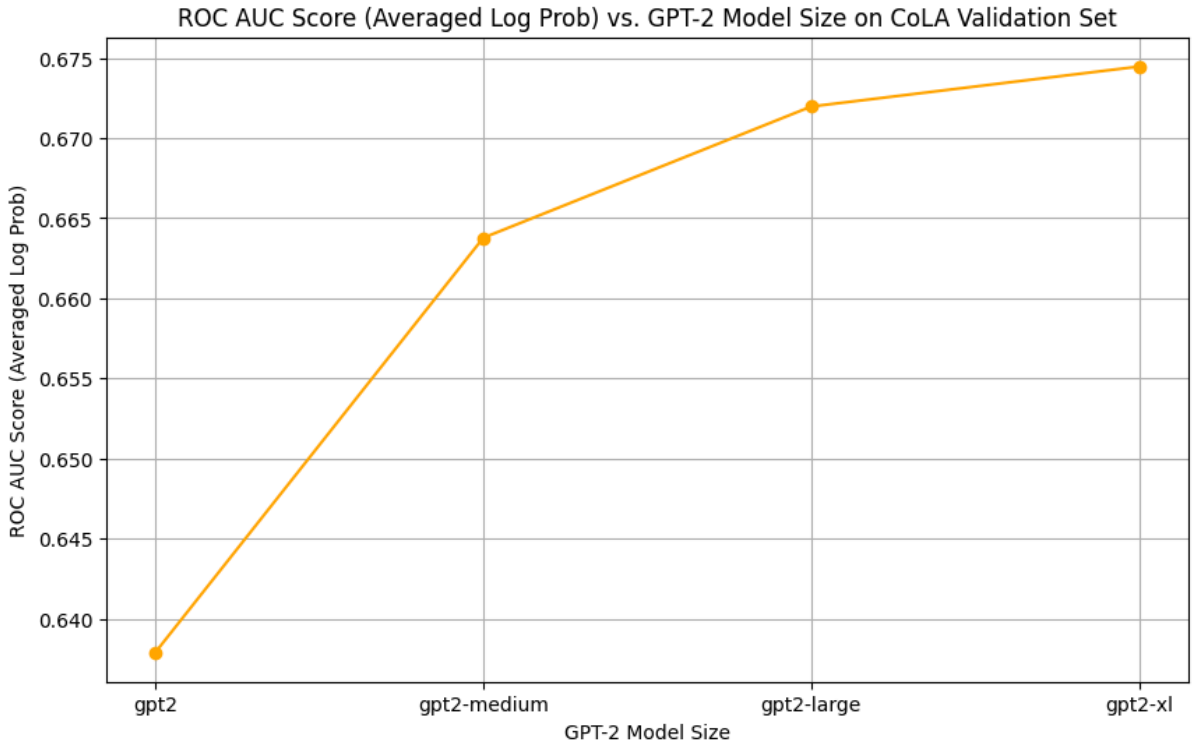
הניתוח שלנו עקב אחר ארבעה גדלים שונים של מודלי GPT-2, הזמינים בפלטפורמת Hugging Face, במטרה לבדוק האם ההסתברויות שמקצה המודל למשפטים מתואמות עם הערכות אנושיות של תקינות דקדוקית.

תחילה חקרנו את הדגם הבסיסי והקטן ביותר של GPT-2. את הלוג'יטס של המודל המרנו להסתברויות באמצעות פונקציית Log-Softmax, ועבור כל מיקום טוקן זיהינו את ההסתברות הלוגריתמית שהוקצתה לטוקן הנכון במשפט הקלט. לאחר מכן חישבנו את ההסתברות הלוגריתמית הכוללת של המשפט, באמצעות סכימת לוג-הסתברויות של כלל הטוקנים התקפים במיקומיהם. לבסוף, הערכנו את הקשר שבין הסתברות המשפט לתקינותו הדקדוקית בעזרת מדד ROC AUC.

מדד ROC AUC (Receiver Operating Characteristic – Area Under Curve) בודק עד כמה המודל מסוגל להבחין בין שתי קטגוריות — במקרה שלנו, בין משפטים תקינים לשאינם תקינים דקדוקית. המדד מתאר את השטח שמתחת לעקומת ה-ROC, כאשר ערך גבוה יותר (קרוב ל-1) מצביע על יכולת הבחנה טובה יותר של המודל, וערך של 0.5 משקף ניחוש אקראי. את אותו התהליך חזרנו וביצענו גם על שלושה גדלים נוספים של המודל (medium, large, xl), ולבסוף השווינו בין תוצאותיהם.



בשלב הבא, חזרנו על כלל שלבי החישוב, אך במקום להשתמש בסכום ההסתברויות הלוגריתמיות, חישבנו את ממוצע הלוג הסתברויות עבור כל משפט על ידי נרמול לפי אורך המשפט (מספר הטוקנים). נרמול זה נועד להפחית את ההטיה הנובעת ממשפטים ארוכים במיוחד, ולהבטיח השוואה הוגנת בין משפטים באורכים משתנים. נמצא כי שיטה זו הניבה ערכים גבוהים יותר של ROC AUC לכל גדלי המודלים.



בהשוואה בין שתי השיטות (Summed לעומת Averaged) נמצא כי ממוצע הלוג הסתברויות הוביל ברוב המקרים לתוצאות ROC AUC גבוהות יותר, דבר המלמד כי נרמול לפי אורך המשפט מפחית השפעות לוואי של שונות באורך ומאפשר הערכה מדויקת יותר של סבירות המשפט כיחידה סמנטית קוהרנטית. הסיבה לכך נעוצה בחישוב: סכימת ההסתברויות הלוגריתמיות מייצרת ערכים הגדלים באופן ישיר עם אורך המשפט, ולכן משפטים ארוכים עלולים לצבור ערכי הסתברות גבוהים יחסית, גם אם אינם תקינים דקדוקית, פשוט בשל ריבוי טוקנים. לעומת זאת, חישוב ממוצע ההסתברויות מנרמל את השפעת מספר הטוקנים, כך שהציון הסופי מבטא טוב יותר את סבירות המשפט כולו כיחידה לשונית אחת, ללא תלות באורכו.

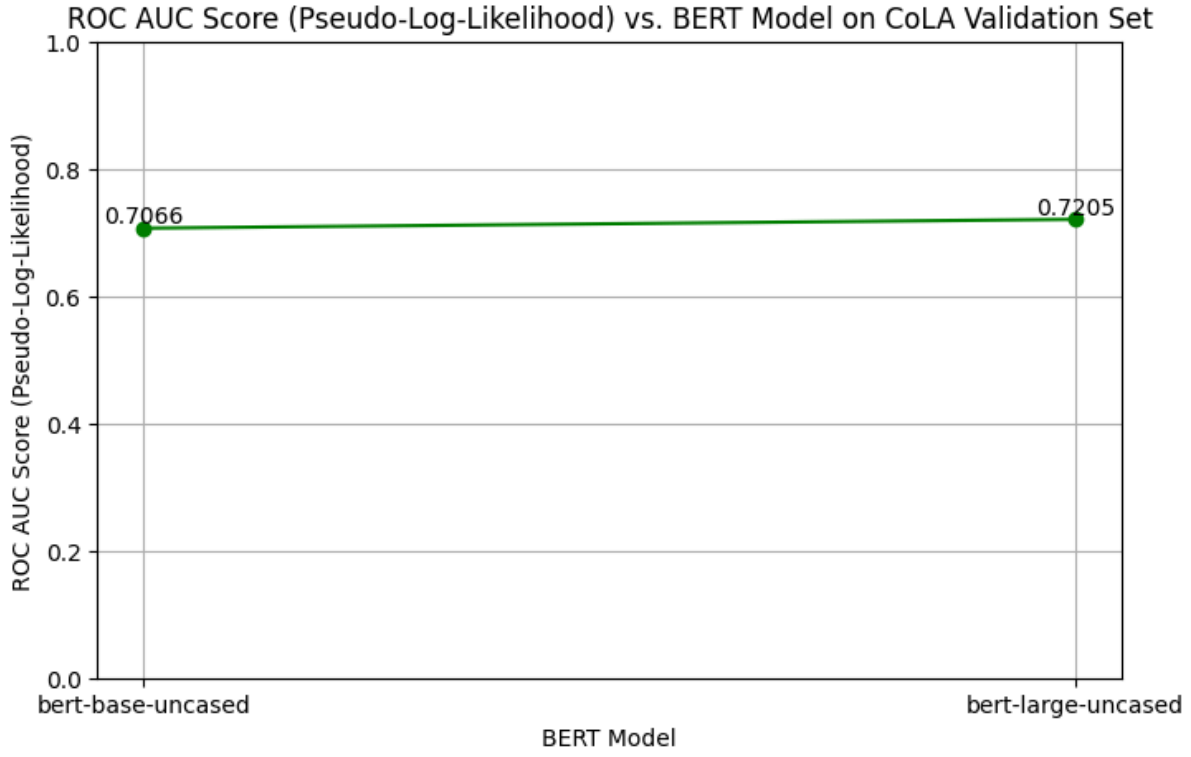
בנוסף, ניתן להסיק כי קיים קשר עקבי וברור בין גודל המודל ממשפחת GPT-2 לבין יכולתו לנבא שיפוטים אנושיים של תקינות דקדוקית, כפי שבא לידי ביטוי בעלייה מתמדת בערכי ה-ROC AUC. העלייה חלה הן כאשר מחשבים את סכום ההסתברויות הלוגריתמיות והן כאשר מחשבים את ממוצען, אך בשיטה השנייה (Averaged Log Prob) מתקבלים ערכי ROC AUC גבוהים יותר באופן עקבי לכל גודל מודל. עוד ניתן לראות כי עיקר השיפור מושג במעבר מ- `gpt2` ל- `gpt2-medium`, בעוד שהשדרוג לגרסאות הגדולות יותר (`gpt2-large` ו-`gpt2-xl`) מביא לשיפורים הדרגתיים ושוליים יחסית.

Part 2: Bidirectional Language Models for Acceptability Prediction

בשלב זה בחנו את יכולתם של מודלים דו-כיווניים (Bidirectional) בעיקר ממשפחת BERT להעריך קבילות דקדוקית של משפטים בשיטת Masked Language Modeling. לצורך כך השתמשנו במודל bert-base-uncased, גם מ- Hugging Face. לכל משפט בסט הולדיציה של CoLa חישבנו Pseudo Log Likelihood (כלומר Best-Score):

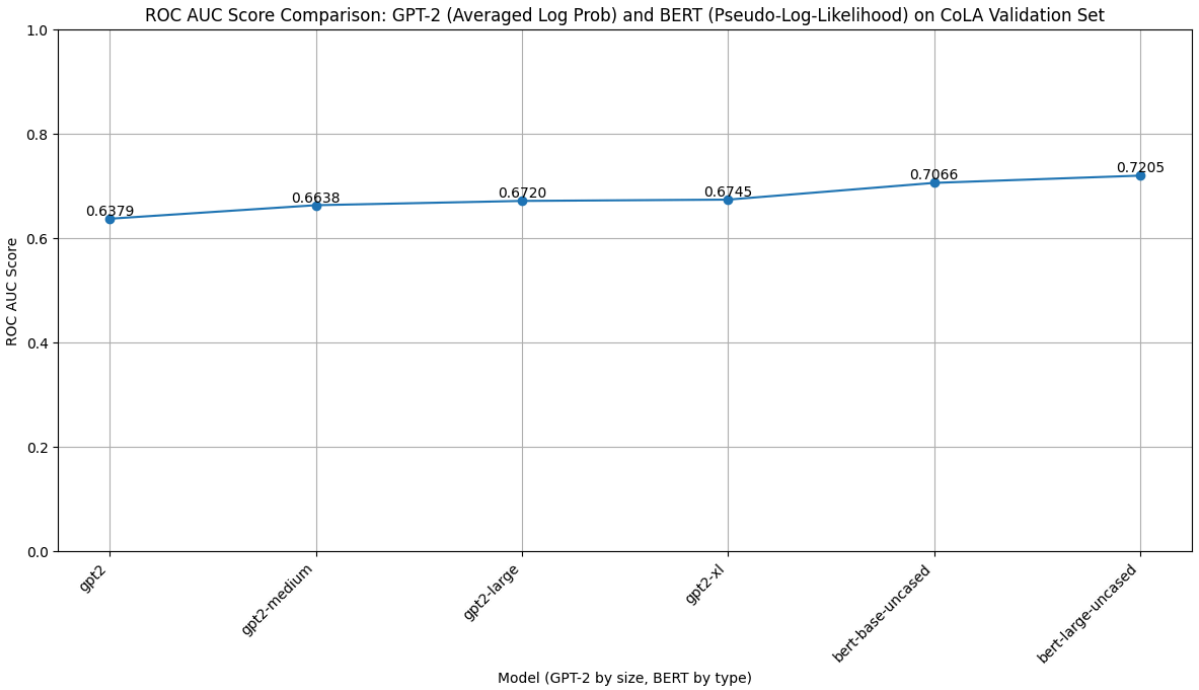
1. הפעלנו לולאת איטרציה על כל עמדות הטוקנים
2. בכל איטרציה הסרנו טוקן בודד (מסקינג) ממיקומו המקורי
3. המודל חזה מחדש את הטוקן המוסתר והחזיר את ההסתברות הלוגריתמית של הטוקן הנכון.
4. סכימה על ערכי ההסתברות האלו יצרה את ציון הbest score לכל משפט.

באופן זה נוצרת אינדיקציה כוללת לסבירותו הדקדוקית של המשפט כפי שמשתקפת מהיכולת של המודל לשחזר כל רכיב תחבירי תוך התחשבות בהקשר המלא. חישבנו ROC AUC לערכי הbest score לעומת השיפוטים האנושיים עבור שני דגמי BERT: הראשון bert base uncased וbert large uncased:



מגרף זה עולה שגם כאן, הגדלת המודל מהבסיס לגדול שיפרה את הביצועים אך באופן שולי, כלומר, כבר הגרסה הבסיסית מראה רמת דיוק גבוהה והשיפור של המודל המתקדם יותר כמעט זניחות, בסוף המטלה אנחנו נראה כיצד לקחנו את אותו מודל גדול יותר והצלחנו לגרום לו להציג הבדלים הרבה יותר משמעותיים.

בהמשך, רצינו לבצע השוואה ישירה בין שתי משפחות המודלים, הן מבחינת גודל והן מבחינת מבנה (אוטורגרסיבי לעומת דו-כיווני), ביחס ליכולתם להעריך קבילות דקדוקית באופן טבעי, ללא Fine-Tuning. יצרנו גרף המציג את ערכי ה־ROC AUC הכי גבוהים שחושבו עבור שתי משפחות המודלים שנועד להמחיש האם הגדלת גודל המודל משפיעה גם במקרה של BERT באופן דומה להשפעה שנצפתה ב־GPT-2.



מניתוח הגרף המשולב עולה בבירור כי משפחת BERT, הפועלת בשיטה דו-כיוונית עם Masking, מפגינה יכולת חיזוי עדיפה על פני משפחת GPT-2 האוטורגרסיבית. ערכי ה־ROC AUC של מודלי BERT גבוהים בעקביות מהערכים שהשיגו המודלים ממשפחת GPT-2, גם כאשר מדובר בגרסאות הגדולות ביותר של GPT-2. אפשר להסיק שקיים יתרון מובנה של BERT בקליטת הקשר לשוני בשני הכיוונים במקביל, שעוזר לו במשימות הדורשות רגישות להקשרים תחביריים וסמנטיים כאחד.

ניתן לקבוע כי מודל BERT מהווה פתרון טוב יותר כפרוקסי Out-of-the-Box להערכת קבילות דקדוקית, גם בלי Fine-Tuning. היתרון של BERT מתבטא ביכולתו להחזיר ציון עקבי ורלוונטי לשיפוטים אנושיים, בעוד שמודלים אוטורגרסיביים כדוגמת GPT-2 אמנם מציגים שיפור עם גידול בגודל, אך נותרים מאחור מבחינת יכולת הקביעה הסופית.

Part 3: Fine-Tuning a Model on CoLA

בשלב האחרון במטלה עשינו Fine-Tuning למודל BERT large uncased. טענו את גרסת Auto Models For Sequence Classification המיועדת לסיווג בינארי של משפטים לפי קבילות.
פיצלנו את סט האימון של CoLa לתת סט אימון ולתת סט ולידציה פנימי כדי שנוכל לנטר את תהליך האימון ולכיול ההיפרפרמטרים. הרצו חמישה אפוקים והערכנו את הדיוק (accuracy) לאחר כל אפוק על סט הולידציה המקורי שמתאים לדיוק במשימות שבהן יש חוסר איזון בין דוגמאות תקינות ללא תקינות.
כך תהליך הfine tuning שלנו נראה:

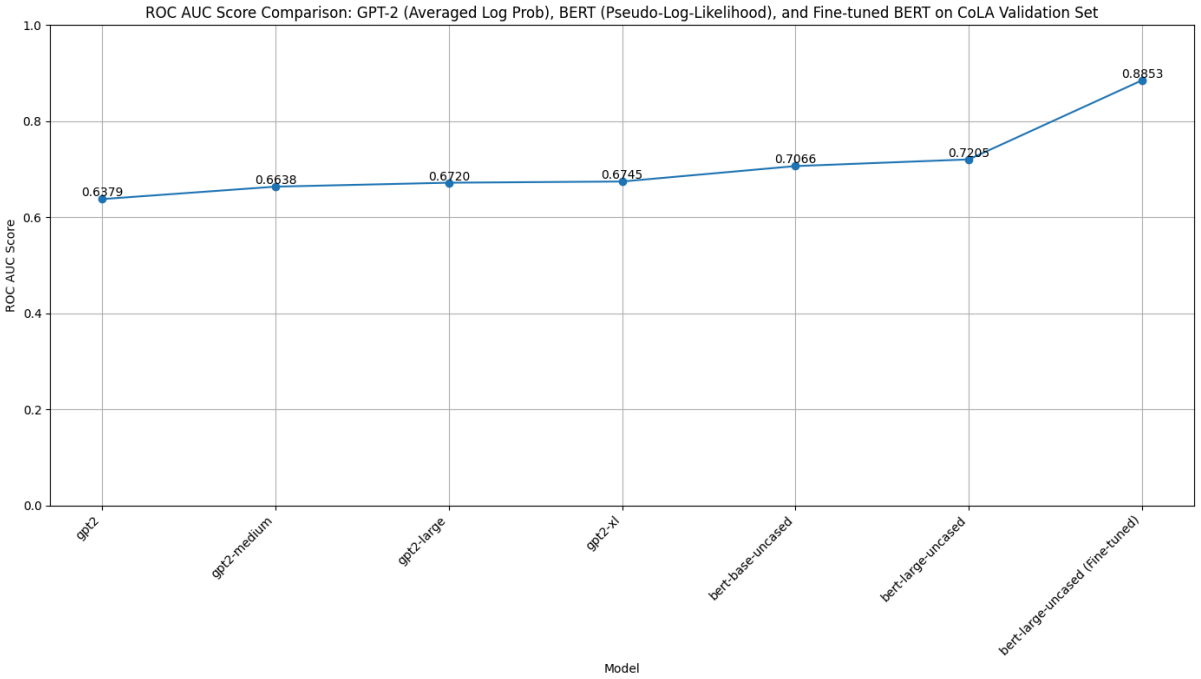
Epoch	Training Loss	Validation Loss	Accuracy
1	No log	0.421318	0.816836
2	No log	0.541512	0.816836
3	0.420600	0.437842	0.841777
4	0.420600	0.513415	0.847233
5	0.158000	0.709619	0.848012

אז כפי שניתן לראות ככל שהמודל עבר על המידע כולו יותר פעמים (epoch גדול יותר) כך שגיאת האימון קטנה ושגיאת הולידציה עלתה אך מצד שני הדיוק המשיך לעלות, מה שמעיד על כך שהמודל משתפר בביצועים שלו אך חששנו כי ישנה אפשרות למידה מסוימת של אוברפיטינג.

לשם כך וידאנו את יכולת ההכללה של המודל בעזרת כך שהרצנו את הפרדיקציות של המודל על סט ולידציה של מידע חדש שהוא לא ראה לפני:

Best model found at step 1140. Approximately corresponds to Epoch 5.
{'eval_loss': 0.7892110347747803, 'eval_model_preparation_time': 0.0093,
'eval_accuracy': 0.8360498561840843, 'eval_runtime': 7.6331, 'eval_samples_per_second': 136.641, 'eval_steps_per_second': 17.162}

הדיוק של המודל על מידע חדש נראה ממש קרוב למידת הדיוק שלו בזמן הfine tuning ולכן ניתן להסיק כי מידת ההכללה שלו היא טובה והדבר האחרון שהחלטנו לעשות איתו זה לחשב את הROC AUC שלו להשוות את הערך הזה שלו למודלים קודמים ולהכניס הכל לגרף אחד:



המסקנה היא חד משמעית, הfine tuning שיפר משמעותית את הדיוק והראה יתרון **מובהק** לעומת השימוש במודל pretrained בלבד.