# Review 150: Language Modeling Is Compression, 21.09.2023

**Paper: https://arxiv.org/abs/2309.10668v2**

https://huggingface.co/papers/2309.10668

מאמר זה משך את ליבי מיד כשראיתי את הכותרת. הרי יש לא מעט קשרים בין מודלי חיזוי לבין מודלי כניסה. מאוד מאוד בגדול מודל חיזוי חזק אמור להפיק את פיצ'רים המהותיים ביותר של הדאטה הנחוצים לחיזוי ואותם הפיצ'רים ניתן לנצל לדחיסת הדאטה.

היום ב-#shorthebrewpapereviews סוקרים מאמר מרתק של דיפמיינד שמראה שניתן להשתמש במודלי שפה מאומנים בתור דוחסי דאטה. איך הם בכלל ניגשו לבעיה הזו הרי זה לא לגמרי טריוויאלי איך ניתן להשתמש במודלי שפה לדחיסת דאטה. הרי מה שמודלי שפה (דקורר) יודעים לעשות הוא לחזות את התפלגות הסתברויות הטוקנים בהינתן הטוקנים הקודמים בסדרה.

אז התברר שקיימת שיטת דחיסה שמנצלת הסתברויות אלו לדחיסת דאטה. השיטה נקראת קידוד אריתמטי(arithmetic encoding). עבור פיסת דאטה נתון השיטה מתחיל מאינטרוול [0,1]. בהגעה של הטוקן הבא מחלקים את האינטרוול לפי ההסתברויות בהתלפגות הטוקן הבא(בהינתן הקודמים) ולוקחים את האינטרוול המתאים לטוקן הבא בדאטה. לאחר הגעת הטוקן האחרון לוקחים את האינטרוול שהתקבל ולוקחים ממנו מספר שניתן לייצגו במספר המינימלי של ביטים.

סדרת ביטים זו תהווה את הדחיסה של פיסת הדאטה שלנו. אז המחברים לקחו את השיטה ובדקו עד כמה ניתן לדחוס דאטהסטים שונים מ 3 דומיינים (טקסט, תמונות, אודיו) ודחסו אותם עם הגישה הזו. לאחר מכן הם השוו את התוצאה עם הדוחסים הקלאסיים כמו gzip, LZMA2 ו- PNG (יש פרטים מעניינים איך המחברים התמודדו עם אורכי הקשר שונים בין מודלי שפה לדוחסים הקלאסיים) ויצא שמודלי שפה מצליחים להביא קצב דחיסה גבוה יותר ב 3 הדומיינים האלו.  
  
אבל תזכרו שיש כאן קאץ' קטן. דוחסים אלו לא לוקחים בחשבון את גודל המודל שנצטרך לשמור אותו אם נרצה לפענח את הדאטה). הדוחסים הקלאסיים הם מאוד קטנים ושם זה פחות רלוונטי. זה למעשה מביא אותנו לקונספט מתמטי מורכב הנקרא סיבוכיות קולמוגורוב שמודד את ״מורכבות״ הדאטה בתור סכום של תוכנה (המודל) לדחיסת הדאטה וגודל הדאטה אחרי הדחיסה. הם מדברים על זה קצת במאמר – ממליץ להעיף מבט.