# Review 206b: [Short] Training LLMs over Neurally Compressed Text

**Paper: https://arxiv.org/abs/2404.03626v3**

https://arxiv.org/pdf/2404.03626.pdf

נתקלתי במאמר החמוד הזה של DeepMind and Anthropic

מה הוא בעצם מציע? לאמן מודל שפה לא על טקסט כמו שאנו רגילים היום אלא על טקסט מקומפרס. זה מגניב כי מודלי שפה ידועים ביכולתם לדחוס טקסט לייצוגים דחוסים אבל זה סיפור טיפה שונה.

אז מה בעצם נותן לנו אימון של llm על טקסטים דחוסים. קודם כל אימון מהר יותר, אורך הקשר ארוך יותר ויש עוד כמה. אז מה הבעיה? זה קצת עדין - הרי אם אנו דוחסים דאטה עם אלגוריתם חזק התוצאה תהיה רעש רנדומלי (אחרת המודל ילמד וינצל את זה).

אז מה המאמר בעצם עשה? הוא לקח מודל שפה M1 שאומן על סדרות ביטים שמייצגות את הטקסט ודחס את הפלט שלו. כמובן M1 גם דוחס את הדאטה (הרי זה מודל שפה) אבל לטענת המחברים בצורה רחוקה ממושלמת. אז הם לקחו שיטת דחיסה קלאסית הנקראת arithmetic coding(AC וכאמור דחסו את הפלט של M1. הם גם יצרו טוקנים חדשים אבל הפעם כל טוקן מיוצג על ידי צ'אנק של ביטים (באורך קבוע) הדוחס את של ביטי הקלט. כאן AC לוקח את ההסתבריות ש-M1 מוציא לכל טוקן ודוחס אותם. לאחר הפיכתם של סדרות אלו לטוקנים "הדחוסים״ מאמנים מודל שפה איתם בצורה הרגילה.

מעניין שניתוח ביצועים נעשה על ידי השוואות של perplexity המנורמל עם מקדם דחיסת דאטה (יודעים למה?). בסך הכל מאמר חמוד.