# Review 79: [Short] SqueezeLLM: Dense-and-Sparse Quantization

**Paper: https://arxiv.org/abs/2306.07629v4**

מה הדרך ההגיונית לשמור ולהריץ מודלי שפה במחשבים אישיים לא חזקים במיוחד? כמובן לשמור את המשקלים שלה לא בדיוק המלא (32 ביט) אלא בדיוק חלקי (נגיד 4 ביט). אך איך לעשות זאת כדי לא לספוג ירידה קשה בביצועים?

הקוונטיזציה היוניפורמית לא תספק לנו כאן ביצועים טובים כי המשקלים לא מפולגים יוניפורמית. המאמר מציע לקלסטר משקלים עם k-means כאשר המרחק מבוסס על החשיבות של המשקלים למודל (כמה הם משפיעים על הלוס). זה מאשפר לשמור משקולות שחשובים יותר בדיוק גבוה כאשר הפחות חשובים נשמרים בדיוק נמוך.

המחברים גם שמו לב כי ה-outlierים משפיעים לרעה על איכות הקלסטרים (מורחים אותם) והחליטו לשמור אותם בקלסטרים נפרדים מהשאר

https://arxiv.org/abs/2306.07629