# Review 133: [Short] OmniQuant: Omnidirectionally Calibrated Quantization for Large Language Models, 30.08.23 https://arxiv.org/abs/2308.13137

**Paper: https://arxiv.org/abs/2308.13137v3**

קוונטיזציה (quantization) של מודלי שפה הפך להיות נושא חם לאחרונה. מודלי שפה הפכו להיות ענקיים ומכילים עשרות מיליארדי פרמטרים וגודלם נמדד בגיגבייטים. כדי לאפשר הפעלה מהירה של מודלי שפה (וגם חסכון במקום האחסון).   
  
היום ב- #shorthebrewpapereviews סוקרים גישה חדשה לקוונטיזציה של מודלי שפה המשלבת טכניקות שהוצעו בכמה עבודות קודמות וגם מציעה כמה חידושים. נתחיל מרענן מהי מטרת הקוונטיזציה? המטרה היא ״לדחוס״ את מודל השפה באופן שהפגיעה בביצועים (inference) תהיה מינימלית (בד״כ בהינתן תקציב דחיסה נתון כלומר מקדם הדחיסה).   
  
קודם כל המחברים מייצגים את בעיית הקוונטיזציה בבעיית מזעור(minimization) של ההפרש בין תוצאת החישוב של המודל ב-precision מלא והתוצאה של המודל המקוונטט (על סט ולידציה) עבור כל בלוק טרנספורמר (יחידה בסיסית של כל llms היום). עכשיו נשאלת השאלה מה הפרמטרים של בעיית מזעור זו?   
  
המאמר מציע לשלב שתי טרנספורמציות נלמדות (פר בלוק הטרנספורמר): הראשונה היא פעולת קוונטיזציה עצמה של משקלי המודל (מתבצעת באמצעות טרנספורמציית min-max עם שני פרמטרים של scaling נלמדים).   
  
הפעולה השניה היא טרנספורמציה לינארית של פלטי הביניים של המודל עם פרמטרים נלמדים. למשל הפלט של בלוק טרנספורמר מסוים מוכפל במטריצה ומוזז (פרמטרים נלמדים) לפני שהוא נכנס לבלוק הטרנספורמר הבא. נציין כי יש פרמטרים שונים לטרנספורמציה לינארית בכניסה למנגנון תשומת הלב. שילוב קווינטוט עם טרנספורמציה של הפלט (משום מה נקרא activation במאמר שזה קצת מבלבל) מביא לדחיסה יעילה של מודל שפה עם פגיעה מינורית בביצועים.