# Review 141: [Short] DOLA: DECODING BY CONTRASTING LAYERS IMPROVES FACTUALITY IN LARGE LANGUAGE MODELS, 08.09.2023 https://huggingface.co/papers/2309.03883

**Paper: https://arxiv.org/abs/2309.03883v2**

אנחנו משתמשים במודלי שפה למשימות רבות אבל האם לסמוך על פלטיהם? עם כל הצער עדיין לא. יש תופעה הנקראת הזיות (hallucinations) של מודלי שפה כאשר מודלי שפה מדברים שטויות. זו בעיה מאוד רצינית בטח אם אתם רוצים לשים מודל שפה בפרודקשן.   
  
אז היום ב- #shorthebrewpapereviews אנחנו סוקרים מאמר המציע פתרון לבעיה החמורה הזו. המחברים מבססים את שיטתם על התכונה המעניינת של מודלי שפה שניתן לצפותה כאשר מוסיפים שכבת סופטמקס המחשבת את התפלגות הטוקנים אחרי כל בלוק הטרנספורמר.   
  
למעשה יש כאן שתי תופעות נפרדות. עבור טוקנים קלים (יחסית) לניחוש (הנובעים מכללי הדקדוק למשל) אז התפלגות טוקנים משכבות האמצע בערך לא משתנה ושווה להתפלגות הסופית של הטוקנים. בטוקנים היותר קשים ההתפלגות משתנה משמעתית כמעט עד השכבה האחרונה – כלומר בשכבה לפני האחרונה התפלגות הטוקנים עשויה להיות שונה מאוד מהשכבה הסופית.   
  
אחד ההסברים לכך הוא ״שואב את ״הידע העובדתי מהזיכרון שלו״ (סליחה על נפנופי הידיים אבל ככה כתוב במאמר). המחברים מציעים לנרמל את הסופטמקס הסופי עם הסופטמקס של השכבה בעלת שוני הגבוה ביותר (מבחינת התפלגות הטוקנים). כלומר הסתברות התוקן פרופורציאונלית להשתנות המקסימלית של ההסתברות של הטוקן הזה (= ״כמות הלמידה״?).   
  
עושים זאת רק לטוקנים בעלי הסתברות גבוהה מספיק בשכבת סופטמקס האחרונה – השכבות עם שוני מקסימלי נבחרות על סט ולידציה. השוני נמדד במונחי Jensen-Shannon Divergence או JSD (גרסה סימטרית של KL) בין התפלגויות הטוקנים. מכיוון שנרמול זה עלול לדפוק את הדקדוק מוסיפים ״קנס על חזרתיות״ (שלא יפלוט את אותו הטקסט יותר מפעם אחת).