⚡️🚀המאמר היומי של מייק 11.07.24: ⚡️🚀

DOLA: DECODING BY CONTRASTING LAYERS IMPROVES FACTUALITY IN LARGE LANGUAGE MODELS

המאמר שנסקור היום הולך להיות די קליל. הוא מתמקד בהקטנת הזיות (hallucinations) של מודלי שפה. מה זה הזיה של מודל שפה? זו שאלה לא טריוויאלית בכלל (יש כמה תרחישים). נתמקד בהזיה המתבטאת בכך שהמודל נותן תשובה לא נכונה עובדתית. נגיד, כלומר על השאלה מה עיר בירה של לטביה הוא עונה שזה ריגה בזמן שהתשובה הנכונה היא טאלין.

המחברים מציע שיטה ה״מכיילת״ את התפלגות הטוקנים בשכבת החיזוי (האחרונה) של מודל שפה. המאמר טוען כי בהרבה מקרים שבהם הטוקנים הנכונים בתשובה מפגינים עליה משמעותית בהסתברות מהשכבות הראשונות ועד האחרונות. זה בולט במיוחד בטוקנים הלא טריוויאלים (לא מילות חיבור וכאלו) הדורשים ממודל שפה לגייס את הידע העובדתי שלו. בהתאם לאובזקבציה זו המאמר מציע שיטה המורכבת משני שלבים. בשלב הראשון מזהים את השכבה הרחוקה ביותר מבחינת התפלגות הטוקנים (השכבה הזו נקראת השכבה הכי פחות בשלה) מהשכבה האחרונה. מרחק כאן מוגדר על ידי Jensen-Shannon divergence או JSD בין התפלגויות הטוקן.

בשלב השני מחסירים (ב-log scale) את ההסתברויות של השכבה הכי פחות בשלה מההסתברויות של השכבה האחרונה. בנוסף מאפסים את כל לוגיטים של הטוקנים בעלי הסתברות הקטנות ביותר (שממילא לא אומורים להיבחר). לאחר מכן עושים סופטמקס ומשתמשים בשיטת decoding האהובה עליהם כדי לחזות את הטוקן הבא.

https://arxiv.org/abs/2309.03883