# Review 120: [Short] SIMPLE SYNTHETIC DATA REDUCES SYCOPHANCY IN LARGE LANGUAGE MODELS, 10.08.23 https://arxiv.org/abs/2308.03958.pdf

**Paper: https://arxiv.org/abs/2308.03958v2**

חנפנות של מודלי שפה, נתקלתם כבר בתופעה הזו? חנפנות או sycophancy באנגלית זו תופעה שבה מודל שפה מושפעים בצורה מופרזת ממה שהמשתמש אומר להם גם אם זה לא נכון באופן אובייקטיבי. למשל אם תגידו למודלי שפה שאתם חושבים ש 1+1=50 ואז תשאלו אותו האם זה נכון בלא מקרים מודל השפה יגיד שכן.   
  
היום ב-#shorthebrewpapereviews סוקרים מאמר שמנסה לגרום למודלי שפה לא להיות חנפן מדי ולהיצמד לעובדות ולא למה שהמשתמש אומר במקרים רלוונטיים. הפתרון שלהם מאוד פשוט – הם בונים דאטהסט של דוגמאות מהסוג הזה ומכיילים מודל שפה עליו. איך הם בנו את הדאטהסט הזה? הם לקחו 17 דאטהסטים מ-huggingface של שאלות-תשובות ובחרו 100K דוגמאות באקראי לכל מודל שפה שהם כיילו (4 גדלים של Flan-PaLM).   
  
לאחר מכן הם הורידו את הדעה של המשתמש ואז השוו את תשובתו של מודל שפה ללא משתנה בעקבות כך. לאחר מכן לוקחים דוגמאות שהמודל ״משתכנע״ בקלות ומשנה את דעתה בעקבות דעת המשתמש ומרכיבים מהם דאטהסט אבל עם התשובה הנכונה. המחברים לפעמים מוסיפים פרטים על המשתמש שעלולים לבלבל מודל שפה (למשל להוסיף שהמשתמש פרופסור למתמטיקה חושב שהפתרון לתרגיל נכון למרות שהוא לו). וזהו זה – מכיילים מודל שפה על הדאטהסט זה.