# Review 155: [Short] CHAIN-OF-VERIFICATION REDUCES HALLUCINATION IN LARGE LANGUAGE MODELS, 27.09.2023

**Paper: https://arxiv.org/abs/2309.11495v2**

https://arxiv.org/abs/2309.09888.pdf

מכירים את בעיית ההזיות (hallucinations)במודלי שפה? בגדול זה קורה כאשר מודל שפה מספק לנו תשובות לא נכונות לשאלות לפעמים יחסית פשוטות. סוגיה זו קיבלה התייחסות רבה לאחרונה במספר עבודות ומאמר שנסקור היום ב-#shorthebrewpapereviews מציע גישה נוספת לפתרונה.

הרעיון של עליו מתבססת השיטה המוצעות במאמר הוא פשוט ומאוד אינטואיטיבי ומסתמך על אובזרבציה הבאה: יש לא מעט מקרים שמודלי שפה מספקים לנו תשובות לא נכונות אך כאשר מבקשים ממנו לבדוק את התשובה הוא חוזר בו ונותן תשובה נכונה (אציין שלא תמיד זה קורה מסיבות טובות אלא נובע מאופן האימון שלו הגורם למודל להיות קצת ״yesman" עם האנשים.   
  
אבל אולי כדאי לנו לבקש מהמודל לבדוק את תשובתו על ידי גנרוט שאלות רלוונטיות לתשובה. למשל אם המודל עונה על שאלה איזה פוליטיקאים ידועים נולדו בשנות השישים, והוא עונה X אז אחת משאלות הבדיקה יכולה להיות "מתי נולד פוליטיקאי X". על סמך תשובות על השאלות אלו (שכמובן יכולת להיות רבות ומגוונות) המודל משנה את תשובתו והופך אותה למדויקת יותר. וזה וזה – כל השאר זה הנדסת פרומפרטים ל- 5 השלבים הבאים:

שאלה מקורית

תשובה התחלתית של מודל שפה

גנרוט שאלות וריפיקציה לשאלה ההתחלתית וגנרוט תשובות עליהם (מתבצע בשני שלבים)ֿ

מודיפיקציה של התשובה הסופית בהתבסס על התוצאה של השלב הקודם.