# Review 124: [Short] Self-Alignment with Instruction Backtranslation, 14.08.23 https://huggingface.co/papers/2308.06259

**Paper: https://arxiv.org/abs/2308.06259v3**

נניח שאתם רוצים לבצע יישור (alignment) עם דאטה מתויג (בסגנון של instruction tuning) של מודל השפה שלכם שאימנתם קודם על דאטה גדול ולא מתויג. נניח שיש בידיכם דאטה מתויג איכותי לא גדול במיוחד ודאטהסט מאוד גדול ומגוון אך לא מתויג.   
  
המאמר שנסקו היום ב-shorthebrewpapereviews מציע שיטה אינטואיטיבית ואלגנטית להפקה של דאטהסט איכותי מתויג בגודל משמעותי מהדאטהסט הלא מתיוג שיש ברשותנו. תהליך האימון מורכב משני שלבים עיקריים: קודם כל מכיילים מודל מאומן ליצור הוראה (instruction) מהתשובה עם הדאטהסט האיכותי המתויג שיש לנו.   
  
לאחר מכן מזינים למודל את ה״תשובות״ מהדאטהסט הלא מתויג כדי ליצור הוראה לכל לכל תשובה. שלב הזה נקרא self-augmentation. כמובן שלא כל הזוגות שיצרנו הם באיכות גבוהה ואנו מפלטרים אותם בשלב השני הנקרא self-curation. לוקחים מודל שמכויל רק עם הדוגמאות מהדאטהסט המתויג האיכותי (הקטן). מבקשים את המודל (עם פרומפט ספציפי) לדרג מ-1 עד 5 את התאמת התשובה להוראה.   
  
לאחר מכן מפלטרים את הזוגות בעלי ציונים הנמוכים. המחברים גם הציעו מה ש נקרא iterative self-curation שבמהלכו לוקחים את הזוגות (הוראה, תשובה) בעלי ציונים גבוהים, מכיילים את המודל עם זה. לאחר מכן ניתן למנף את המודל ל-self-augmentation (השלב הראשון) כדי ליצור דאטהסט מתויג איכותי עוד יותר. ניתן לחזור על התהליך כמה פעמים בתקווה לקבל דאטהסט מתויג גדול ואיכותי. מאמר אלגנטי ונחמד…