# Review 119: [Short] FLIRT: Feedback Loop In-context Red Teaming, 09.08.23 https://huggingface.co/papers/2308.04265

**Paper: https://arxiv.org/abs/2308.04265v2**

בטיחות מודלים גנרטיביים הינו אחד מנושאי המחקר החמים בבינה מלאכותית גנרטיבית (GenAI). הרי אנחנו לא רוצים מודל המצייר תמונה לפי התיאור הטקסטואלי יגנרט לנו תמונה קשה, אלימה או מטרידה גם אם נבקש את זה ממנו. למניעת תופעות אלו צריך לזהות פרומפטים מתוחכמים שגורמים למודל ליצור תוכן בעייתי.   
  
היום ב-#shorthebrewpapereviews סוקרים מאמר המציע גישה לזיהוי פרומפטים זדוניים שעלולים לגרום ליצירת תוכן מסוכן. המאמר מציע לבנות סטים של פרומפטים זדוניים הממקסמים 3 מטריקות שכל אחת מהם מודדת היבט שונה של ״זדוניות״ הפרומפטים מהסט הזה. היעד הראשון הוא מקסום סבירות של יצירת תוכן מסוכן עם פרומפטים מהסט, השני הוא הגיוון הסמנטי של הפרומפטים (כמה שפחות דמיון בין הפרומפטים) והיעד השלישי הוא הנראות ״הטובה״ של פרומפטים אלו (כלומר העדר של מילים גסות או בעלות תוכן מיני מובהק).  
  
המאמר משתמש במודל שפה בשביל ליצור פרומפטים אלו באמצעות מנגנון למידה in-context. האלגוריתם מתחיל בכמה פרומפטים זדוניים שנכתבו על ידי בני אדם ואז משתמשים במודל שפה כדי לגנרט פרומפטים זדוניים באמצעות מודל שפה (למידה in-context). עבור כל פרומפט זדוני שהצליח (יצר תוכן מסוכן) יוצרים סטים שבהם כל פרומפט מהסט מוחלף בפרומפט החדש ובוחרים מהם את הסט שממקסם לנו פונקציית היעד.  
  
איך יודעים שפרומפט הצליח ליצור תוכן לא ראוי? משתמשים במודלים מאומנים לזיהוי תוכן לא בטוח (כמו NudeNet או Q16).