# Review 178: [Short] CAPSFUSION: Rethinking Image-Text Data at Scale

**Paper: https://arxiv.org/abs/2310.20550v3**

https://huggingface.co/papers/2310.19019

הרבה דובר לאחרונה על איך ניתן לשפר מודלים מולטימודליים (אלו שיודעים לעבוד עם כמה סוגים של דאטה, נגיד שפה ותמונות). רוב מודלים אלו אומנו על שדאטהסטים גדולים מהאינטרנט המכילים תמונות והכותרת שלהם. המודל הראשון המפורסם שאימן מודל כזה היה כמובן CLIP.

מכיוון שמודלים אלו מאומנים עם כמות עצומה (עשרות מיליונים של תמונות או יותר) של דאטה בדיקה של איכות הדאטה אינה אפשרית. אז איך ניתן לשפר את איכות ההתאמה בין התמונה לכותרת שלה ללא שימוש בבודקים אנושיים. ניחשתם נכון – ניתן להשתמש ב-ChatGPT.

איך עושים זאת? לוקחים מודל שמאומן לתת כותרות לתמונות ויוצרים כותרת לתמונה איתו. לאחר מכן לוקחים את הכותרת המקורית ומבקשים מ- ChatGPT למזג אותן יחד לכותרת אחת עם איזה פרומפט מהונדס היטב. ד״א הפרומפט מכיל הוראה המונעת מ-ChatGPT מיזוג של שתי הכותרות. באמצעות פעולה פשוטה שכזו איכות הכותרת ויישורה לתמונה משתפרת מאוד. בסוף לוקחים את הדאטה עם הכותרות החדשות ומכיילים על זה מודל מולטימודאלי. לפי התוצאות במאמר עובד לא רע.