# Review 99: [Short] Scaling Autoregressive Multi-Modal Models: Pretraining and Instruction Tuning

**Paper: https://arxiv.org/abs/2309.02591v1**

המאמר הפופולרי של השבוע נסקר היום ב- #shorthebrewpapereviews:

כמובן שזה CM3Leon/Chameleon של @Meta

המחברים אימנו מודל שיודע לעבוד עם תמונות וגם עם טקסט. אבל כבר ראינו מודלים כאלו (Stable StableDiffusion, DaLLE2), ImageBind שיודע לעבוד עם 6 סוגים של דאטה. אז מה מיוחד ב-CM3Leon?

2 דברים מבליטים את CM3Leon מהמודלים האחרים שיודעים שעובדים עם סוגי דאטה שונים:

הוא דו-כיווני: הופך טקסט לתמונה וגם תמונה לטקסט

המודל אומן בהתחלה כמודל שפה טהור כאשר התמונות (עם התיאור) שימשו רק לאחזור, כלומר אימון retrieval-augmented . בשלב השני עשו לו כיול עם שילוב ישיר של תמונות

בשלב השני פשוט מחפשים את התמונה עם התיאור הקרוב ביותר לתיאור טקסטואלי נתון ואז משנים אותה לפי השוני בין התיאורים של שתי תמונות.

המודל משדרג את ארכיטקטורה שהוצעה לראשונה ב -RA-CM3 שהוא בעצמו שכלול של CM3 המפורסם (יחסית). מה שמעניין כאן שגם התמונה וגם הטקסט מחולקים לטוקנים: VQ-GAN.

https://arxiv.org/abs/2309.02591