# Review 167: [Short] Reward-Augmented Decoding: Efficient Controlled Text Generation With a Unidirectional Reward Model

**Paper: https://arxiv.org/abs/2310.09520v4**

https://huggingface.co/papers/2310.09520

מי שעקב אחרי המהפכות שהתרחשו בעולם של מודלי שפה ענקיים (LLMs) בטח שמעו על RLHF שזה ראשי תיבות של Reinforcement Learning with Human Feedback (או בקצרה RLHF) בהקשר של אימון מודלי שפה.

המאמר שנסקור היום ב-#shorthebrewpapereviews לוקח אחת מאבני הבניין של RLHF שזה מודל תגמול (reward model) ומשתמש בה לגנרוט של טקסט. מודל תגמול מיועד לשערוך של איכות הטקסט המגונרט על ידי המודל ומרטת RLHF היא למקסם את התגמול (יחד עם עוד כמה מדדים) במטרה לשפר את איכות הטקסט המגונרט. המאמר המסוקר משתמש למודל התגמול לגנרוט של טקסט בפרט ל"כיול״ של הסתברויות של הטוקנים שמודל שפה מחשב בשביל לחזות כל טוקן.

כלומר עבור כל טוקן נחזה הסתברותו מוזזת בהתאם לתגמול המצופה על ידי הוספת טוקן זה לטוקנים שכבר גונרטו על ידי המודל. טוקנים בעלי הסתברות גבוהה לפי מודל השפה וגם בעלי ערך גבוה של פונקציית התגמול (הממודלת על ידי מודל תגמול) יקבלו עדיפות על פני הטוקנים בעלי ערכי התגמול נמוכים יותר.

מודל התגמול מאומן התאם למשימה נתונה עם פונקציית לוס של המחשבת מרחק בין את התגמול ה-ground truth לזה של המודל לכל טוקן. מעניין כי ככל הקנס על תגמול לא מדויק עולה ככל שהטוקן רחוק יותר מהתחלת הטקסט המגונרט (הקנס על אי דיוק של הטוקן האחרון הוא מקסימלי).