המאמר היומי של יניב ומייק: 09.06.25  
Spurious Rewards: Rethinking Training Signals in RLVR – Fast Overview

המסר המרכזי במשפט אחד  
 גם תגמולים אקראיים או שגויים יכולים להביא לשיפור דרמטי ביכולות פתרון בעיות מתמטיות – אבל רק אם המודל כבר "מכיר" את הדרך מהפרה-טריינינג.

### למה זה חשוב

למידה באמצעות חיזוקים עם תגמול ניתן לאימות (RL with Verifiable Rewards - RLVR) הפכה לשיטה מובילה לשפר יכולות חשיבה של מודלים גדולים. המאמר שואל שאלה פרובוקטיבית: האם אנחנו באמת צריכים תגמול מדויק? התשובה: לא תמיד.

### מה עשו החוקרים

הם לקחו את המודל Qwen-2.5-Math ואימנו אותו על סט שאלות מתמטיקה עם חמש גרסאות שונות של תגמולים:

תגמול אמיתי: מודל מקבל נקודה רק אם התשובה נכונה.

תגמול לפי הצבעת רוב: המודל מייצר 64 תשובות, ומתגמל את התשובה השכיחה.  
תגמול פורמטי: אם התשובה כוללת ביטוי מתמטי (למשל \boxed{}), היא מתוגמלת, בלי קשר לנכונות.

תגמול אקראי: הטלת מטבע קובעת אם לתגמל.

תגמול הפוך: רק תשובות שגויות מקבלות נקודה.

במפתיע, כל אחד מהתגמולים הללו הצליח כמעט כמו תגמול אמיתי כלומר המודל השתפר דרמטית גם כש האות החיזוקי לא היה קשור כלל לתוצאה הנכונה.

### ממצאים עיקריים

Qwen משתפר בכל תנאי: גם בלי תגמול נכון, המודל לומד לפתור בעיות טוב יותר. לעומת זאת, מודלים אחרים (כמו Llama3 ו־OLMo2) זקוקים לתגמול מדויק כדי להשתפר.

הגורם הסמוי: פתרון דרך קוד. Qwen כבר יודע לנסח פתרונות בפייתון מתוך הטקסט. אימון RLVR רק גורם לו לבחור באסטרטגיה הזו יותר ומביא לדיוק גבוה יותר.

שיפור בדיוק נובע ממעבר מ"לשוני" ל"קוד": בשאלות שבהן המודל התחיל לכתוב קוד בעקבות האימון, הדיוק קפץ בכמעט 26%.

אז למה תגמול אקראי עובד? האלגוריתם GRPO כולל קליפינג שמעדיף פעולות בסבירות גבוהה – כך שגם כשאין קשר לתוצאה, המודל לומד לחזק את ההתנהגות הדומיננטית שלו.

לא כל מודל נולד שווה: כשאין במודל נטייה מוקדמת לקוד, כמו ב־OLMo, אותו תגמול אקראי פשוט לא עובד.

סיכום:

המאמר מראה שלעיתים קרובות אימון RL לא מלמד כישורים חדשים, אלא מחלץ כישורים חבויים שהמודל כבר פיתח בפרה-טריינינג. לא תמיד צריך תגמול מדויק – אם המודל כבר "מכיר" את הדרך, מספיק לאותת לו לחזור אליה. עם זאת, זה לא נכון לכל מודל – יש כאלה שדורשים הנחיה מדויקת כדי להשתפר.

https://arxiv.org/abs/2412.07169