SimPO: Simple Preference Optimization with a Reference-Free Reward

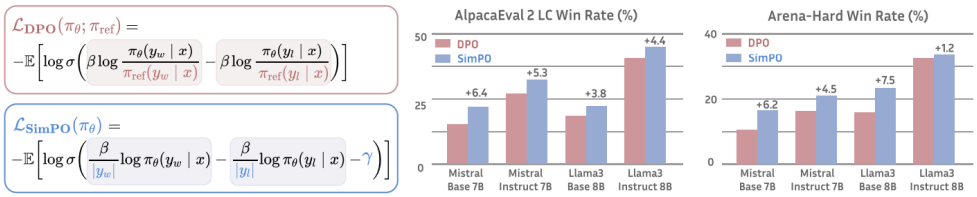

Figure 1: SimPO and DPO mainly differ in their reward formulation, as indicated in the shaded box. SimPO outperforms DPO across a wide range of settings on AlpacaEval 2 and Arena-Hard.

המאמר שנסקור דן בנושא אימון של מודלי שפה. אתם בטח יודעים שאימון מודל שפה foundational מורכב מ- 3 שלבים עיקריים:

1. אימון מודל self-supervised על דאטהסט ענק
2. אימון(פיינטיון) מפוקח (supervised fine-tuning או SFT) על דאטהסט מתויג קטן יותר (בד״כ מכיל תשובות רצויות למגוון שאלות) במטרה לגרום למודל לעקוב אחרי הוראות המשתמש (instruction following)
3. שלב RLHF: מתברר שרוב המודלים לא מצליחים ללמוד רק מהתשובות ה״טובות״ ואנו נדרשים לספק לו גם את התשובות ה״לא טובות״. השלב האחרון נעשה באמצעות שימוש בטכניקות השונות של למידה עם חיזוקים.

המודלים הראשוניים (גוגל, OpenAI) שהשתמשו ב-RLHF ליישור (alignment) של המודלים התבססו על טכניקה שנקראת Proximal Policy Approximation או PPO בקצרה. במהלך האימון אנו מעדכנים את המודל שלנו כך שהוא ייתן תגמול (reward=) גבוה לתשובה טובה לתשובה ותגמול נמוך לתשובה לא טובה תוך שמירה של המודל החדש קרוב (מבחינת התפלגויות הטוקנים) שהוא מוציא להתפלגות המתקבלת בשלב 2.

אבל איך נמדוד את התגמול הזה? עבור PPO אנו צריכים לאמן מודל תגמול שבהינתן פרומפט ותשובה יחזיר לנו ציון (סקלרי). עבור תשובה טובה הציון יהיה גבוה ועבור תשובה לא טובה הוא יהיה נמוך. מאמנים את המודל הזה על הדאטהסט של התשובות הטובות ולא טובות משלב 3.

כמובן שאם היה אפשר להסתדר ללא מודל תגמול מצבנו היה טוב יותר. קודם כל זה חוסך לנו את זמן ומשאבים ובנוסף אנו לא צריכים להפעיל אותו לאינפרנס במהלך אימון RLHF שזה גם יכול להפחית את דרישות הזכרון וכוח חישוב. אז הוצעו שיטות כמו Direct Preference Optimization או DPO שהוריד את הצורך באימון מודל תגמול. לאחר מכן יצא מודל הנקרא ORPO (סקרנו אותו באנגלית לפני כחודש) הסתדר גם בלי להשתמש במודל משלב 2 במהלך האימון (משמש רק לאתחול המודל משלב 3).

עכשיו הגענו למאמר המסוקר. הוא הציע שכלול ל-DPO הנקרא SimPo. כמו OrPo הוא לא צריך מודל רפרנס בצורה מפורשת במהלך אימון שלב 3 ומציע לאמן את המודל על ידי מקסום ההפרש בין התגמול של התשובה הטובה והתשובה הלא טובה (עם הסיגמויד) עם איזשהו מרג'ין מסוים. החידוש העיקרי של המאמר שבתור פונקציית תגמול המחברים לוקחים את הנראות המירבית של תשובה בהינתן שאלה, **המנורמלת באורך התשובה** (בטוקנים). המחברים טוענים שדבר זה (נרמול) בין השאר מונע מהמודל לגנרט תשובות ארוכות מדי וזה אכן נשמע די הגיוני.

דרך אגב בנוגע להמרג'ין נטען המאמר שמספר עבודות קודמות ציינו שזה מיטיב עם תהליך האימון (למרות שזה די הוספת קבוע).

"The margin between two classes is known to influence the generalization capabilities of classifiers [1, 9, 19, 27]. In standard training settings with random model initialization, increasing the target margin typically improves generalization".

יש גם את הטבלה החמודה הזו המסכמת את רוב המחקרים האחרונים בתחום RLHF למודלי שפה.

Table 3: Various preference optimization objectives given preference data $\mathcal{D} = (x, y_w, y_l)$, where x is an input, and y_w and y_l are the winning and losing responses.

Method	**Objective**
DPO [62]	$-\log \sigma \left(\beta \log \frac{\pi_\theta(y_w\|x)}{\pi_{\text{ref}}(y_w\|x)} - \beta \log \frac{\pi_\theta(y_l\|x)}{\pi_{\text{ref}}(y_l\|x)}\right)$
IPO [6]	$\left(\log \frac{\pi_\theta(y_w\|x)}{\pi_{\text{ref}}(y_w\|x)} - \log \frac{\pi_\theta(y_l\|x)}{\pi_{\text{ref}}(y_l\|x)} - \frac{1}{2\tau}\right)^2$
KTO [25]	$-\lambda_w \sigma \left(\beta \log \frac{\pi_\theta(y_w\|x)}{\pi_{\text{ref}}(y_w\|x)} - z_{\text{ref}}\right) + \lambda_l \sigma \left(z_{\text{ref}} - \beta \log \frac{\pi_\theta(y_l\|x)}{\pi_{\text{ref}}(y_l\|x)}\right)$, where $z_{\text{ref}} = \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\beta \mathrm{KL}\left(\pi_\theta(y\|x)\|\pi_{\text{ref}}(y\|x)\right)\right]$
ORPO [38]	$-\log p_\theta(y_w\|x) - \lambda \log \sigma \left(\log \frac{p_\theta(y_w\|x)}{1-p_\theta(y_w\|x)} - \log \frac{p_\theta(y_l\|x)}{1-p_\theta(y_l\|x)}\right)$, where $p_\theta(y\|x) = \exp\left(\frac{1}{\|y\|} \log \pi_\theta(y\|x)\right)$
R-DPO [60]	$-\log \sigma \left(\beta \log \frac{\pi_\theta(y_w\|x)}{\pi_{\text{ref}}(y_w\|x)} - \beta \log \frac{\pi_\theta(y_l\|x)}{\pi_{\text{ref}}(y_l\|x)} - (\alpha\|y_w\| - \alpha\|y_l\|)\right)$
SimPO	$-\log \sigma \left(\frac{\beta}{\|y_w\|} \log \pi_\theta(y_w\|x) - \frac{\beta}{\|y_l\|} \log \pi_\theta(y_l\|x) - \gamma\right)$