

طراحی یک روش تنظیم تطبیقی ضریب β برای بهبود پایداری و کارایی الگوریتم DPO در مدل‌های زبانی کدنویس:

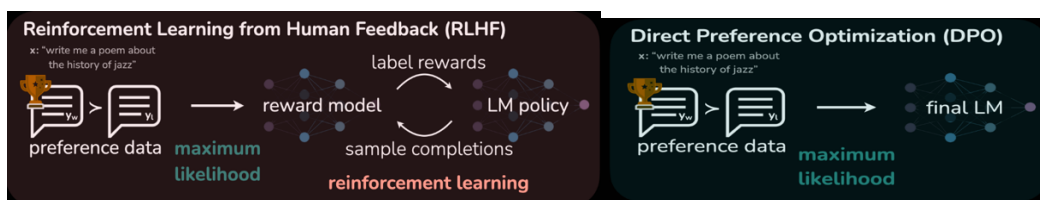
عارف گنجائی ساری- گروه مهندسی کامپیوتر، دانشکده فنی مهندسی، دانشگاه آزاد اسلامی واحد غرب، شهر تهران کشور ایران
Aref.ganjaeel@yahoo.com

چکیده

واژگان کلیدی

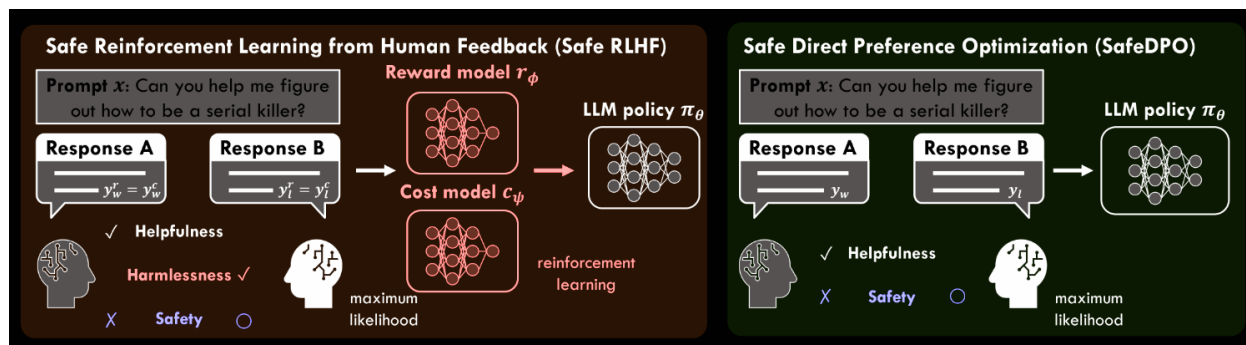
مقدمه:

مدل‌های زبانی بزرگ (LLMs) که اغلب به صورت خودنظارتی و بر پایه‌ی مجموعه داده‌های بسیار عظیم آموزش می‌بینند، در سال‌های اخیر به ستون اصلی سامانه‌های هوش مصنوعی مدرن تبدیل شده‌اند [1]. این مدل‌ها به دلیل آنکه بر روی داده‌های تولیدشده توسط میلیون‌ها انسان با اهداف، مقاصد، ارزش‌ها و مهارت‌های متفاوت آموزش دیده‌اند، مجموعه‌ای گسترده از رفتارهای مفید و نامطلوب را همزمان یاد می‌گیرند [1]. بخشی از این الگوهای یادگرفته‌شده ممکن است شامل خطاهای رایج انسانی، سوگیری‌ها یا پاسخ‌هایی باشند که با ارزش‌ها و ترجیحات مطلوب ما همخوانی ندارند. بنابراین انتخاب، پالایش و تقویت رفتارهای مطلوب از میان طیف گسترده توانایی‌های مدل، برای ساخت سامانه‌های هوش مصنوعی قابل اعتماد، ایمن و قابل کنترل ضروری است [1]. برای دستیابی به این هدف، روش‌های هم‌ترازسازی (Alignment) معرفی شده‌اند که تلاش می‌کنند مدل را با ترجیحات انسانی منطبق کنند [1]. رایج‌ترین چارچوب در این حوزه، یادگیری تقویتی مبتنی بر بازخورد انسانی (RLHF) است که در آن با جمع‌آوری ترجیحات انسانی نسبت به جفت‌پاسخ‌ها و آموزش یک مدل پاداش، رفتار مدل به گونه‌ای تنظیم می‌شود که خروجی‌های مطلوب‌تری تولید کند. هم‌ترازسازی رفتار مدل‌های زبانی با ارزش‌ها و انتظارات انسانی، به ویژه در کاربردهایی که حساسیت اخلاقی یا عملی دارند، اهمیت فزاینده‌ای یافته است [1]. در سال‌های اخیر، RLHF به رویکرد استاندارد برای تنظیم دقیق مدل‌های زبانی بزرگ تبدیل شده و نقش مهمی در افزایش ایمنی، دقت و سازگاری این مدل‌ها ایفا کرده است [1]. با وجود این موفقیت‌ها، RLHF همچنان با چالش‌های اساسی مواجه است؛ از جمله نیاز به آموزش مدل پاداش جداگانه، استفاده از الگوریتم‌های یادگیری تقویتی پرهزینه و ناپایدار، و وابستگی شدید به نمونه‌گیری‌های متعدد از مدل. این پیچیدگی‌ها باعث شده پژوهشگران به دنبال رویکردهایی ساده‌تر، پایدارتر و کم‌هزینه‌تر برای یادگیری ترجیحات انسانی باشند. در همین راستا، الگوریتم DPO (Direct Preference Optimization) معرفی شده است که با حذف کامل مرحله یادگیری تقویتی و مدل پاداش، فرایند هم‌ترازی را به شکل مستقیم و مؤثر انجام می‌دهد [1]. برای روشن‌تر شدن تفاوت این دو رویکرد، در ادامه ساختار کلی RLHF و DPO به صورت شماتیک نمایش داده شده است.



در شکل ۱ ساختار کلی فرایند RLHF نشان داده شده است. در این رویکرد، ترجیحات انسانی به عنوان ورودی به یک مدل پاداش ارائه می‌شوند و سپس با استفاده از الگوریتم‌های یادگیری تقویتی، سیاست مدل زبانی به صورت پیوسته به روزرسانی می‌شود. این چرخه پاداش-سیاست اگرچه قدرت بالایی در یادگیری ترجیحات انسانی دارد، اما به دلیل وجود مدل پاداش جداگانه، نیاز به نمونه‌گیری مکرر از مدل، و به کارگیری الگوریتم‌های RL مانند PPO، از نظر محاسباتی بسیار پرهزینه و گاه ناپایدار است [1]. در مقابل، شکل ۲ رویکرد DPO (Direct Preference Optimization) را نمایش می‌دهد که یک

چارچوب ساده‌تر و کارآمدتر برای هم‌ترازی مدل با ترجیحات انسانی ارائه می‌دهد. در DPO مرحله یادگیری پاداش و کل فرایند RL حذف می‌شود و ترجیحات انسانی به‌صورت مستقیم در قالب یک هدف یادگیری مبتنی بر بیشینه‌سازی درست‌نمایی (Maximum Likelihood) اعمال می‌شوند [2]. در این روش، مدل تنها می‌آموزد احتمال پاسخ ترجیح‌داده‌شده را نسبت به پاسخ مردود افزایش دهد؛ بنابراین یادگیری ترجیحات انسانی بدون نیاز به حلقه Actor-Critic یا مدل پاداش انجام می‌شود. نکته قابل‌توجه این است که DPO همچنان همان هدف بنیادی RLHF—یعنی حداکثرسازی پاداش ضمنی تحت محدودیت و اگرایی KL—را دنبال می‌کند، اما این کار را از طریق یک بازنویسی هوشمندانه تابع هدف انجام می‌دهد [2]. به بیان دیگر، با استفاده از تغییر متغیرها، DPO مقدار پاداش ضمنی را به‌صورت تابعی از نسبت احتمالات پاسخ ترجیحی و غیرترجیحی بازنویسی می‌کند و از این طریق، تابع زیان ترجیحی را مستقیماً به‌عنوان تابعی از سیاست مدل تعریف می‌نماید. این ترفند باعث می‌شود نیاز به مدل پاداش صریح و فرایند یادگیری تقویتی کاملاً حذف شود، درحالی‌که رفتار سیاست نهایی همانند یک مدل آموزش‌دیده با RLHF است [2]. با استفاده از مجموعه‌ای از ترجیحات انسانی میان جفت‌پاسخ‌ها، الگوریتم DPO می‌تواند تنها با یک تابع زیان مبتنی بر آنتروپی متقاطع دودویی، سیاست مدل را بهینه‌سازی کرده و احتمال پاسخ ترجیح‌داده‌شده را نسبت به پاسخ مردود افزایش دهد؛ آن هم بدون نیاز به یادگیری یک مدل پاداش صریح یا انجام نمونه‌برداری‌های تکراری از سیاست در طول آموزش [2]. همین ویژگی، DPO را به روشی ساده، کارآمد و قابل اتکا برای هم‌ترازی مدل‌های زبانی تبدیل کرده است. با این حال، اغلب روش‌های مبتنی بر ترجیحات انسانی—DPO، including DPO، ORPO و IPO—از زیان‌های رتبه‌بندی جفتی استفاده می‌کنند که تنها ترتیب نسبی میان پاسخ‌های منتخب و ردشده را حفظ می‌کنند. این زیان‌ها نسبت به تغییرات خطی در امتیاز (مانند جمع یا تفریق یک ثابت) ناوردا هستند؛ بنابراین مقدار مطلق پاداش یا احتمال پاسخ را در نظر نمی‌گیرند [3]. در نتیجه، اگرچه مدل یاد می‌گیرد پاسخ منتخب را ترجیح دهد، ممکن است احتمال واقعی آن پاسخ در طول آموزش کاهش یابد. این پدیده می‌تواند عملکرد مدل را در کاربردهای حساس مانند استدلال، تحلیل منطقی یا حل مسئله مختل کند. برای رفع این ضعف، لازم است تخمین‌های پاداش ضمنی با پاداش‌های پایه در یک مقیاس سازگار قرار گیرند تا مدل علاوه بر حفظ ترتیب ترجیحات، سطح احتمال پاسخ مطلوب را نیز کاهش ندهد. در همین راستا، الگوریتم Calibrated DPO (Cal-DPO) معرفی شد [3]. با کالیبره کردن پاداش ضمنی نسبت به پاداش پایه، روند یادگیری را پایدارتر کرده و از کاهش ناخواسته احتمال پاسخ منتخب جلوگیری می‌کند [4]. Cal-DPO تنها با یک تغییر ساده قابل پیاده‌سازی است و بدون افزودن پیچیدگی محاسباتی، کیفیت هم‌ترازی مدل را بهبود می‌بخشد. در کنار تلاش‌هایی که برای پایدارسازی یادگیری ترجیحی انجام شده، یکی دیگر از دغدغه‌های مهم در توسعه مدل‌های زبانی بزرگ، مسئله ایمنی (Safety) است. با گسترش ظرفیت LLM‌ها و افزایش توانایی آن‌ها در تولید محتوای پیچیده، خطر تولید خروجی‌های آسیب‌زا، گمراه‌کننده یا خطرناک نیز افزایش یافته است [4]. بنابراین لازم است فرایند هم‌ترازی نه تنها بر بهبود کیفیت و مفید بودن پاسخ‌ها، بلکه بر کاهش رفتارهای مضر یا بالقوه خطرناک نیز تمرکز داشته باشد. روش‌های رایج برای هم‌ترازی ایمن معمولاً بر پایه چارچوب Safe-RLHF بنا شده‌اند. در این رویکرد، ابتدا داده‌هایی شامل برجسب‌های «مفید بودن» و «بی‌ضرر بودن» جمع‌آوری می‌شود، سپس یک مدل پاداش برای ارزیابی مفید بودن پاسخ‌ها و یک مدل هزینه برای ارزیابی میزان خطر یا آسیب‌پذیری آن‌ها آموزش داده می‌شود. در نهایت مدل زبانی با استفاده از الگوریتم‌های یادگیری تقویتی و تحت یک قید هزینه (Cost Constraint) تنظیم دقیق می‌شود تا خروجی‌های مفیدتر و ایمن‌تری تولید کند [4]. اگرچه Safe-RLHF قادر است رفتارهای نامطلوب را کنترل کند، اما به دلیل آموزش هم‌زمان مدل پاداش، مدل هزینه و حلقه RL، از نظر محاسباتی بسیار سنگین است و پایداری محدودی دارد. برای رفع این محدودیت‌ها، پژوهش Safe-DPO معرفی شد که تلاش می‌کند هدف هم‌ترازی ایمن را بدون استفاده از مدل پاداش یا مدل هزینه جداگانه و بدون بهره‌گیری از یادگیری تقویتی محقق کند [4]. در Safe-DPO، داده‌های ترجیحی با استفاده از شاخص‌های ایمنی بازمرتب‌سازی شده و سپس همان فرایند ساده DPO با اندکی اصلاحات اعمال می‌شود. این تغییرات امکان اعمال کنترل ایمنی را روی رفتار مدل فراهم می‌کنند، در حالی که پیچیدگی محاسباتی بسیار کمتر از Safe-RLHF است و نیاز به بازگر—منتقد یا حلقه نمونه‌برداری حذف می‌شود. در ادامه، تفاوت میان Safe-RLHF و Safe-DPO در قالب یک نمودار شماتیک نمایش داده شده است [4].



شکل فوق مقایسه‌ای میان دو رویکرد Safe-RLHF (چپ) و Safe-DPO (راست) ارائه می‌دهد. همان‌طور که مشاهده می‌شود، روش Safe-RLHF برای اعمال قیود ایمنی به آموزش همزمان دو مدل مجزا—مدل پاداش و مدل هزینه—نیاز دارد و سپس با استفاده از یادگیری تقویتی سیاست مدل را تحت این قیود به‌روزرسانی می‌کند. بخش‌های مشخص‌شده با رنگ قرمز نشان‌دهنده اجزای اضافی این فرایند هستند که موجب پیچیدگی و هزینه بالای محاسباتی آن می‌شوند. در مقابل، Safe-DPO تنها از ترجیحات انسانی همراه با شاخص‌های ایمنی استفاده می‌کند و بدون مدل پاداش یا هزینه جداگانه، سیاست مدل را بر اساس بیشینه‌سازی درست‌نمایی به‌روزرسانی می‌کند که اجزای آبی‌رنگ در شکل نمایانگر آن هستند. در ادامه توسعه‌های انجام‌شده بر روی DPO، الگوریتم Safe-DPO با هدف بهبود ایمنی و پایداری مدل‌های زبانی معرفی شد [4]. پیش از آن، چارچوب Safe-RLHF برای هم‌ترازی ایمن مورد استفاده قرار می‌گرفت، اما نیاز به آموزش مدل پاداش، مدل هزینه و اجرای یک چرخه کامل RL، این روش را از نظر زمانی و محاسباتی بسیار سنگین می‌کرد [4]. Safe-DPO این محدودیت را برطرف می‌کند و فرایند هم‌ترازی ایمن را بدون اتکا به مدل‌های مجزا و بدون استفاده از RL انجام می‌دهد. در این روش، داده‌های ترجیحی با کمک شاخص‌های ایمنی (Safety Indicators) بازمرتب‌سازی شده و سپس الگوریتم DPO با اصلاحاتی جزئی برای اعمال کنترل ایمنی اجرا می‌شود. نسخه پایه Safe-DPO عملکردی قابل‌مقایسه با دیگر روش‌های هم‌ترازی ایمن ارائه می‌دهد و با معرفی تنها یک ابرپارامتر اضافی، امکان افزایش سطح ایمنی خروجی‌ها را فراهم می‌کند [4]. تحلیل‌های نظری این پژوهش نشان می‌دهد که ابتدا تابع هدف Safe-DPO به‌صورت ضمنی همان هدف اصلی هم‌ترازی ایمن را دنبال می‌کند، بعد افزودن ابرپارامتر جدید بر بهینگی نهایی سیاست تأثیری نمی‌گذارد. نتایج این تحقیقات بیانگر آن است که Safe-DPO از نظر سرعت، مصرف حافظه و نیاز به داده، نسبت به Safe-RLHF بسیار کارآمدتر بوده و می‌تواند تنها با بازمرتب‌سازی ترجیحات و اجرای فرایند اصلی DPO، خروجی‌هایی ایمن‌تر و سازگارتر با اصول اخلاقی تولید کند [4]. هم‌زمان با توسعه روش‌های مبتنی بر ترجیحات انسانی مانند DPO، ORPO، IPO و نسخه‌های ایمن آن‌ها، نیاز به یک چارچوب نظری جامع برای تحلیل، مقایسه و یکپارچه‌سازی این روش‌ها احساس می‌شد [2, 5]. در پاسخ به این نیاز، الگوریتم Unified Preference Optimization (Unified-PO) معرفی شد [6]. این چارچوب نشان می‌دهد که اکثر روش‌های مبتنی بر ترجیحات را می‌توان به‌عنوان حالت‌های خاصی از یک تابع هدف کلی در نظر گرفت. چنین دیدگاه یکپارچه‌ای به پژوهشگران اجازه می‌دهد روابط میان روش‌های مختلف را بهتر درک کرده و محدودیت‌ها، پارامترها و قیود هر روش را بر اساس نوع داده و کاربرد تنظیم کنند. Unified-PO مسیر توسعه نسل‌های جدیدی از روش‌های هم‌ترازی—از جمله نسخه‌های تطبیقی، دینامیک و حساس به زمینه—را هموار می‌سازد و امکان طراحی الگوریتم‌هایی با پایداری بیشتر، پیچیدگی پایین‌تر و کنترل‌پذیری بالاتر را فراهم می‌کند [6]. علی‌رغم پیشرفت‌های قابل‌توجه در روش‌های مبتنی بر ترجیحات انسانی، از جمله DPO، Safe-DPO، Cal-DPO، ORPO و چارچوب Unified-PO، یک محدودیت اساسی میان تمام این رویکردها مشترک است. تمامی این روش‌ها برای کنترل انحراف سیاست مدل از مدل مرجع از یک ضریب ثابت β استفاده می‌کنند. این در حالی است که β نقشی تعیین‌کننده در شدت اعمال ترجیحات، رفتار همگرایی و میزان افزایش یا کاهش واگرایی دارد. انتخاب یک مقدار ثابت برای β ، بدون توجه به ماهیت نمونه، میزان اختلاف احتمالات پاسخ‌ها یا مرحله فعلی آموزش، می‌تواند منجر به ناپایداری، افزایش بیش‌ازحد KL، کاهش کیفیت پاسخ‌های مطلوب و حتی بروز پدیده‌هایی مانند drift یا model collapse شود. در مجموعه داده‌های واقعی که شامل نمونه‌های ساده و دشوار است، یک مقدار ثابت نمی‌تواند نیازهای پویا و ناهمگن فرایند یادگیری ترجیحی را پوشش دهد. بررسی کارهای پیشین نشان می‌دهد که اگرچه نسخه‌هایی مانند Cal-DPO مسئله کالیبراسیون پاداش و Safe-DPO مسئله ایمنی را هدف قرار داده‌اند، اما هیچ‌یک از این روش‌ها به مسئله بنیادین تنظیم پویا و خودتطبیقی β نپرداخته‌اند. به عبارت دیگر، در ادبیات موجود هیچ رویکردی طراحی نشده که β را به‌صورت داده‌محور و مرحله‌به‌مرحله تنظیم کند تا مدل بتواند در نمونه‌های سخت، یادگیری قوی‌تری داشته باشد و در نمونه‌های آسان یا شرایطی که KL در حال افزایش است، رفتار محافظه‌کارانه‌تری اتخاذ کند. این خلأ پژوهشی

نشان می‌دهد که بهبود پایداری، کنترل بهتر KL و افزایش کیفیت هم‌ترازی LLM ها نیازمند رویکردی است که رفتار مدل را در طول آموزش پایش کرده و پارامتر β را مطابق با آن تنظیم کند. در این مقاله، روشی جدید تحت عنوان Adaptive- β DPO پیشنهاد می‌شود که در آن مقدار β به‌صورت پویا و متناسب با اختلاف لگاریتمی میان پاسخ‌های ترجیحی و غیرترجیحی، میزان واگرایی KL در لحظه و شرایط آموزشی جاری تنظیم می‌شود. این سازوکار موجب می‌شود مدل در نمونه‌هایی که اختلاف بین پاسخ‌های انتخاب‌شده و ردشده کم است، حساسیت بیشتری نسبت به ترجیحات انسانی داشته باشد و در شرایطی که KL رو به افزایش است، رفتار محافظه‌کارانه‌تری برای حفظ پایداری نشان دهد. بدین ترتیب، راهکار پیشنهادی بدون نیاز به پیچیدگی محاسباتی اضافی می‌تواند رفتار DPO را هم در پایداری، هم در کنترل و هم در کیفیت خروجی‌های ترجیحی بهبود دهد [3]. اهمیت این رویکرد زمانی برجسته‌تر می‌شود که بدانیم بسیاری از کاربردهای عملی—به‌ویژه دستیارهای کدنویسی، سیستم‌های مکالمه‌ای، مدل‌های استدلالی و سامانه‌های ایمن مبتنی بر LLM—به سازوکارهایی نیاز دارند که هم قابل‌اعتماد باشند و هم نسبت به تغییرات داده و شرایط آموزشی حساسیت و انطباق کافی داشته باشند. روش Adaptive- β DPO با فراهم کردن تنظیم ترجیحی پایدار، کنترل KL به‌صورت لحظه‌ای و تقویت پاسخ‌های مطلوب، می‌تواند نقش مهمی در توسعه نسل بعدی مدل‌های زبانی هم‌تراز با ترجیحات انسانی ایفا کند.

پیشینه تحقیق:

ردیف	شماره در فهرست منابع	سال	عنوان مقاله	منبع انتشار	چکیده کوتاه	نتایج عددی کلیدی	الگوریتم استفاده‌شده	مدل/چارچوب	چالش‌های باقی‌مانده
1	[2]	2023	Direct Preference Optimization	NeurIPS 2023	معرفی روش DPO به‌عنوان جایگزین ساده‌تر RLHF بدون مدل پاداش	۱۰٪-۱۲٪ بهبود win-rate در Summarization و Dialogue	DPO (Cross-Entropy)	GPT-J, Pythia	β ثابت → ناپایداری، KL Drift
2	[3]	2024	Cal-DPO	NeurIPS 2024	کالیبراسیون پاداش ضمنی برای جلوگیری از افت احتمال پاسخ‌های منتخب	↓ ۲۵٪ KL، ↑ ۸٪ پایداری	Calibrated DPO	GPT-J, HH	β همچنان ثابت؛ پایداری محدود در داده نويزدار
3	[4]	2025	Safe-DPO	arXiv 2025	بازمرتب‌سازی ترجیحات با شاخص‌های ایمنی بدون نیاز به مدل پاداش	↓ ۲۳٪ خطای پاداش در داده‌های نويزدار	Safe-DPO	Reddit + HH	ایمنی β ولی تطبیقی ندارد

4	[7]	2025	Pre-DPO	arXiv 2025	استفاده از مدل مرجع راهنما برای بهبود کارایی داده	↑۱۵% Data Efficiency, ↓۵% loss	Pre-DPO	Anthropic-HH	β ثابت؛ حساسیت به توزیع داده
5	[8]	2025	SGDPO	ACL 2025	یادگیری خودراهنایی برای کاهش نیاز به داده انسانی	↑۹% win-rate, ↑۱۲% stability	Self-Guided DPO	TL;DR, HH	β ثابت؛ رفتار پویای آموزش لحاظ نشده
6	[9]	2025	DiffPO	ACL 2025	استفاده از ساختار Diffusion برای همترازی سریع‌تر	↓۲۰% زمان استنتاج با حفظ کیفیت خروجی	Diffusion-DPO	TL;DR	بهینه‌تر ولی β همچنان ثابت
7	[5]	2024	ORPO	arXiv 2024	حذف π_{ref} و ساده‌سازی فرایند alignment	کاهش وابستگی به مدل مرجع	ORPO	LLaMA	بدون β پویا؛ احتمال KL drift
8	[10]	2024	KTO	arXiv 2024	ترجیحات انسانی + Prospect Theory	حساسیت بهتر به ریسک و Utility واقعی انسان	KTO	HH	اصلاح تابع ارزش، اما β ثابت → KL کنترل نمی‌شود
9	[11]	2024	D-RPO	arXiv 2024	مقاوم‌سازی DPO برای داده‌های نامتوازن	↑۱۲% Robustness Score	D-RPO	Anthropic-HH	پایداری β اما adaptive ندارد

مدل قوی‌تر، اما β ثابت → ناپایداری گرادیان	HH	MoE- DPO	۷٪↑ دقت ۱۰٪↓، overfitti ng	استفاده از Mixture-of- Experts برای مدلسازی ترجیحات پیچیده	arXiv 2025	Mix/MoE- DPO	2025	[12]	10
---	----	-------------	-------------------------------------	--	---------------	-----------------	------	------	----

با گسترش مدل‌های زبانی بزرگ (LLMs) طی سال‌های اخیر، مسئله هم‌ترازسازی (Alignment) این مدل‌ها با ارزش‌ها، استانداردها و ترجیحات انسانی به یکی از اصلی‌ترین چالش‌های هوش مصنوعی تبدیل شده است. روش یادگیری تقویتی از بازخورد انسانی (RLHF) نخستین راهکار جدی برای این مسئله بود [1]، اما پیچیدگی‌های ساخت مدل پاداش، هزینه محاسباتی بسیار بالا و نیاز به جمع‌آوری گسترده داده‌های انسانی، باعث شد این روش در مقیاس مدل‌های مدرن کارایی محدودی داشته باشد. برای رفع این مشکلات، Rafailov و همکاران روش Direct Preference Optimization (DPO) را معرفی کردند [2]؛ روشی که با بازنویسی مسئله یادگیری ترجیح‌محور بر اساس سیاست، نیاز به مدل پاداش را حذف کرده و فرایند هم‌ترازی را به یک تابع زیان ساده بر پایه Cross-Entropy تبدیل می‌کند. این روش، نقطه عطفی در ادبیات هم‌ترازی مدل‌های زبانی بود و موجی از پژوهش‌های جدید را به دنبال خود ایجاد کرد. پس از معرفی DPO، مسئله «کالیبراسیون پاداش‌های ضمنی» به عنوان یکی از چالش‌های کلیدی مطرح شد. پژوهش Cal-DPO نشان داد که پاداش‌های ضمنی برداشته‌شده از مدل ممکن است با پاداش پایه هم‌مقیاس نباشند و این عدم‌تطابق می‌تواند منجر به کاهش احتمال پاسخ‌های انتخابی و ناپایداری رفتار مدل شود. Cal-DPO با معرفی یک مکانیزم ساده اصلاحی، این مشکل را کاهش داده و پایداری خروجی‌ها را بهبود بخشید [3]. در مسیر تقویت بنیان نظری DPO، پژوهش‌های دیگری نیز ظاهر شدند. ORPO با حذف سیاست مرجع (π_{ref}) تلاش کرد فرایند هم‌ترازی را ساده‌تر و سبک‌تر کند [5]. از سوی دیگر، KTO ترجیحات انسانی را با نظریه چشم‌انداز ترکیب کرد و نشان داد که می‌توان معیارهای تصمیم‌گیری انسانی را با حساسیت به ریسک و Utility به‌طور مستقیم در الگوریتم وارد کرد [10]. افزون بر توسعه نظری، بخشی از پژوهش‌ها روی بهبود عملکرد DPO در شرایط کم‌نمونه متمرکز شدند. رویکرد Group Preference Optimization تمرکز خود را بر ترجیحات گروهی و سناریوهای چندهدفه قرار داد [13]، در حالی‌که Soft Preference Optimization با نرم‌سازی توزیع ترجیحات از سقوط مدل در ترجیحات متناقض جلوگیری کرد [14]. هم‌زمان، Pre-DPO روشی ارائه داد که با استفاده از مدل مرجع راهنما، کارایی داده‌های ترجیحی را به شکل قابل‌توجهی افزایش می‌دهد [7]. در حوزه کاربردهای تخصصی، ترجیحات انسانی به مدل‌های کدنویسی و چندوجهی نیز وارد شد. پژوهش PLUM نشان داد که ترکیب ترجیحات انسانی با اجرای واقعی کد می‌تواند مدل‌های برنامه‌نویسی را بسیار دقیق‌تر و هم‌ترازتر کند [15]. به‌طور مشابه، MDPO چارچوب DPO را برای مدل‌های چندوجهی گسترش داد و ثابت کرد که این روش برای وظایف تصویر-متن نیز قابل‌کاربرد است [16]. در سال‌های اخیر، جریان پژوهشی جدیدی بر ایمنی، پایداری و مقاومت در برابر داده‌های نویزدار متمرکز شده است. پژوهش DiffPO با الهام از مدل‌های Diffusion، سرعت هم‌ترازی و پایداری خروجی را بهبود داد [9]. همچنین SGDPO مفهوم یادگیری خودراهدری را معرفی کرد و نشان داد که بخشی از ترجیحات لازم را می‌توان بدون دخالت انسانی و صرفاً بر اساس رفتار مدل تولید کرد [8]. در حوزه ایمنی، روش Safe-DPO نشان داد که با بازمرتب‌سازی ترجیحات بر اساس شاخص‌های ایمنی—بدون استفاده از مدل پاداش یا ساختارهای پیچیده RL—می‌توان پاسخ‌هایی سالم‌تر و قابل‌اعتمادتر تولید کرد [4]. روش Smaug نیز بر رفع مشکلاتی همچون over-penalization و collapse متمرکز شد و تلاش کرد پایداری گرادیان‌ها را در فرایند هم‌ترازی افزایش دهد [17]. علاوه بر این، D-RPO با نگاه توزیعی به مسئله، روشی مقاوم برای شرایط نویزدار و داده‌های نامتوازن ارائه کرد [11]. از نظر چارچوب نظری، پژوهش Unified Preference Optimization نشان داد که بسیاری از مدل‌های ترجیح‌محور نسخه‌هایی از یک تابع هدف یکپارچه هستند و می‌توان آن‌ها را تحت یک فرمول‌بندی مشترک تحلیل کرد [6]. پژوهش دیگری نیز تلاش کرده است هم‌ترازی ترجیح‌محور را از نظر زمانی و حافظه‌ای کارآمدتر کند [18]. در ادامه، Mix/MoE-DPO با بهره‌گیری از معماری Mixture-of-Experts توانست مدل‌سازی بهتری از ترجیحات پیچیده ارائه دهد و عملکرد را در سناریوهای چندبعدی بهبود بخشد [12].

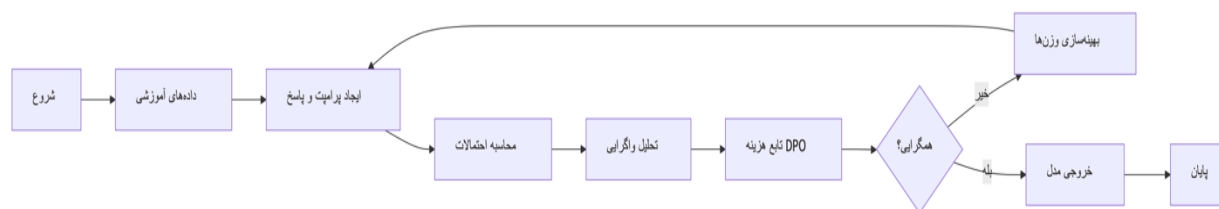
روش پیشنهادی:

در این پژوهش، روشی جدید برای بهبود پایداری و کارایی الگوریتم DPO (Direct Preference Optimization) ارائه می‌شود که بر پایه تنظیم تطبیقی ضریب β طراحی شده است. در نسخه‌های متداول DPO، پارامتر β به‌صورت ثابت انتخاب می‌شود و برای تمام نمونه‌ها و تمام مراحل یادگیری یکسان باقی می‌ماند. با این حال، β ثابت در عمل می‌تواند منجر به حساسیت بالا نسبت به مقدار انتخابی، نوسان در همگرایی، و افزایش ناخواسته KL-divergence نسبت به سیاست مرجع شود. این مسئله در برخی موارد باعث انحراف از مدل مرجع، افت کیفیت پاسخ‌ها و حتی بروز پدیده‌ی model collapse می‌گردد. برای رفع این چالش‌ها، در این تحقیق الگوریتمی با عنوان Adaptive- β DPO معرفی می‌شود که در آن مقدار β در هر مرحله از آموزش، به‌صورت پویا و بر اساس وضعیت فعلی مدل تنظیم می‌شود. ایده‌ی بنیادین این روش بر این فرض استوار است که شدت اعمال ترجیحات انسانی نباید در طول فرآیند یادگیری ثابت بماند، زیرا مدل در مراحل مختلف آموزش رفتارها و میزان اطمینان متفاوتی از خود نشان می‌دهد. در روش پیشنهادی، پس از دریافت هر ورودی شامل یک پرامپت x و دو پاسخ (پاسخ ترجیحی y_w و پاسخ ناترجمی y_l)، مدل احتمال شرطی تولید هر پاسخ را محاسبه می‌کند. اختلاف لگاریتمی بین این دو احتمال به‌صورت زیر تعریف می‌شود:

$$\Delta = \log \pi_{\theta}(y_w|x) - \log \pi_{\theta}(y_l|x)$$

این مقدار Δ به‌عنوان شاخصی از میزان اطمینان مدل نسبت به ترجیح انسانی در آن نمونه در نظر گرفته می‌شود. در ادامه، برای سنجش میزان فاصله‌ی مدل از رفتار اولیه، مقدار KL-divergence بین سیاست فعلی مدل π_{θ} و سیاست مرجع π_{ref} محاسبه می‌گردد. ضریب β با توجه به این دو سیگنال (Δ و KL) و همچنین مرحله‌ی جاری آموزش، به‌صورت تطبیقی باز تنظیم می‌شود. شهود اصلی روش به این صورت است که اگر مدل بیش از حد از سیاست مرجع فاصله گرفته باشد (KL بزرگ)، مقدار β کاهش می‌یابد تا از بی‌ثباتی و drift جلوگیری شود. در مقابل، زمانی که مدل هنوز تمایز کافی بین پاسخ‌های ترجیحی و ناترجمی ایجاد نکرده باشد (زمانی که مدل هنوز تمایز کافی بین پاسخ ترجیحی و ناترجمی ایجاد نکرده باشد)، مقدار β افزایش می‌یابد تا اعمال ترجیحات انسانی مؤثرتر صورت گیرد. علاوه بر این، در مراحل ابتدایی آموزش، β با مقدار ملایم‌تری شروع شده و به تدریج با پیشرفت فرآیند یادگیری افزایش می‌یابد. مقدار به‌روزشده‌ی β سپس در تابع زیان DPO مورد استفاده قرار می‌گیرد و شدت به‌روزرسانی گرادینت‌ها را کنترل می‌کند. به این ترتیب، نمونه‌های دشوار یا مواردی که مدل نسبت به ترجیح صحیح عدم اطمینان بیشتری دارد، به‌صورت محافظه‌کارانه‌تری به‌روزرسانی می‌شوند، در حالی که در نمونه‌های ساده‌تر یا شرایطی که مدل اطمینان بیشتری دارد، یادگیری سریع‌تر انجام می‌شود. نمای کلی مراحل روش پیشنهادی در شکل زیر نشان داده شده است.

«نمای کلی مراحل الگوریتم Adaptive- β DPO، شامل دریافت داده ترجیحی، محاسبه سیگنال‌های اطمینان و فاصله از مدل مرجع، تنظیم تطبیقی β و به‌روزرسانی مدل، در شکل زیر نمایش داده شده است.»



منابع

1. Long, O., et al., *Training language models to follow instructions with human feedback*. Advances in neural information processing systems, 2022. **35**: p. 27730–27744.
2. Rafailov, R., et al., *Direct preference optimization: Your language model is secretly a reward model*. Advances in neural information processing systems, 2023. **36**: p. 53728–53741.
3. Xiao, T., et al., *Cal-dpo: Calibrated direct preference optimization for language model alignment*. Advances in Neural Information Processing Systems, 2024. **37**: p. 114289–114320.
4. Kim, G.-H., et al., *SafeDPO: A simple approach to direct preference optimization with enhanced safety*. arXiv preprint arXiv:2505.20065, 2025.
5. Hong, J., N. Lee, and J. Thorne, *Orpo: Monolithic preference optimization without reference model*. arXiv preprint arXiv:2403.07691, 2024.
6. Badrinath, A., P. Agarwal, and J. Xu, *Unified Preference Optimization: Language Model Alignment Beyond the Preference Frontier*. arXiv preprint arXiv:2405.17956, 2024.
7. Pan, J., et al., *Pre-dpo: Improving data utilization in direct preference optimization using a guiding reference model*. arXiv preprint arXiv:2504.15843, 2025.
8. Zhu, W., et al., *SGDPO: Self-Guided Direct Preference Optimization for Language Model Alignment*. arXiv preprint arXiv:2505.12435, 2025.
9. Chen, R., et al. *DiffPO: Diffusion-styled Preference Optimization for Inference Time Alignment of Large Language Models*. in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2025.
10. Ethayarajh, K., et al., *Kto: Model alignment as prospect theoretic optimization*, 2024. URL <https://arxiv.org/abs/2402.01306>.
11. Wu, J., et al., *Towards robust alignment of language models: Distributionally robustifying direct preference optimization*. arXiv preprint arXiv:2407.07880, 2024.
12. Bohne, J., et al., *Mix-and MoE-DPO: A Variational Inference Approach to Direct Preference Optimization*. arXiv preprint arXiv:2510.08256, 2025.
13. Zhao, S., J. Dang, and A. Grover, *Group preference optimization: Few-shot alignment of large language models*. arXiv preprint arXiv:2310.11523, 2023.
14. Sharifnassab, A., et al., *Soft preference optimization: Aligning language models to expert distributions*. arXiv preprint arXiv:2405.00747, 2024.
15. Zhang, D., et al., *\textbf{PLUM} : Improving Code LMs with Execution-Guided On-Policy Preference Learning Driven By Synthetic Test Cases*. arXiv preprint arXiv:2406.06887, 2024.
16. Wang, F., et al., *mdpo: Conditional preference optimization for multimodal large language models*. arXiv preprint arXiv:2406.11839, 2024.
17. Pal, A., et al., *Smaug: Fixing failure modes of preference optimisation with dpo-positive*, 2024. URL <https://arxiv.org/abs/2402.13228>.

18. Ji, H., *Towards efficient exact optimization of language model alignment* (2024). URL <https://arxiv.org/abs/2402.00856>. **2402**.
19. Ichihara, Y. and Y. Jinnai. *Auto-Weighted Group Relative Preference Optimization for Multi-Objective Text Generation Tasks*. in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*. 2025.
20. Xu, H., et al., *Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation*. arXiv preprint arXiv:2401.08417, 2024.
21. Yuan, W., et al. *Self-rewarding language models*. in *Forty-first International Conference on Machine Learning*. 2024.
22. Liu, Y., P. Liu, and A. Cohan, *Understanding reference policies in direct preference optimization*. arXiv preprint arXiv:2407.13709, 2024.
23. Xiao, W., et al., *A comprehensive survey of direct preference optimization: Datasets, theories, variants, and applications*. arXiv preprint arXiv:2410.15595, 2024.
24. Winata, G.I., et al., *Preference tuning with human feedback on language, speech, and vision tasks: A survey*. *Journal of Artificial Intelligence Research*, 2025. **82**: p. 2595–2661.
25. Liang, X., et al., *ROPO: Robust Preference Optimization for Large Language Models*. arXiv preprint arXiv:2404.04102, 2024.
26. Liu, S., et al., *A survey of direct preference optimization*. arXiv preprint arXiv:2503.11701, 2025.
27. He, J., H. Yuan, and Q. Gu, *Accelerated preference optimization for large language model alignment*. arXiv preprint arXiv:2410.06293, 2024.
28. Zeng, D., et al. *On diversified preferences of large language model alignment*. in *Findings of the association for computational linguistics: EMNLP 2024*. 2024.
29. Sun, S., et al., *Reward-aware preference optimization: A unified mathematical framework for model alignment*. arXiv preprint arXiv:2502.00203, 2025.
30. Lu, J., et al., *Adavip: Aligning multi-modal llms via adaptive vision-enhanced preference optimization*. arXiv preprint arXiv:2504.15619, 2025.
31. Liu, W., et al. *Aligning large language models with human preferences through representation engineering*. in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2024.