بهبود پایداری و کارایی الگوریتم بهینه سازی ترجیحات مستقیم (DPO) از طریق تنظیم تطبیقی ضریب بتا در مدل های زبانی بزرگ برای دستیار های هوشمند کد نویسی:

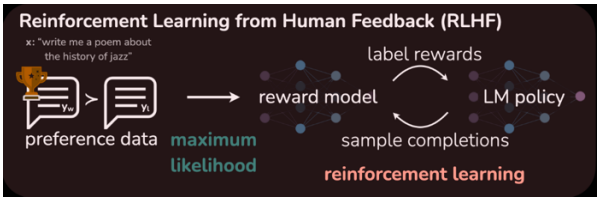
عارف گنجائی ساری- گروه مهندسی کامپیوتر، دانشکده فنی مهندسی، دانشگاه آزاد اسلامی واحد غرب، شهر تهران کشور ایران
Aref.ganjaee1@yahoo.com

چکیده

واژگان کلیدی

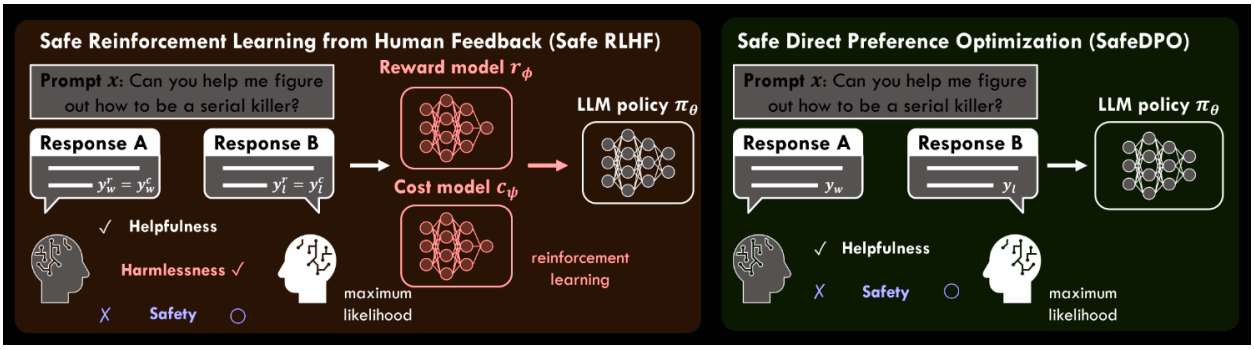
مقدمه:

مدل‌های زبانی بزرگ (LLMs) که اغلب به‌صورت خودنظارتی و بر پایه‌ی مجموعه‌داده‌های 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 قابل‌اعتماد، ایمن و قابل‌کنترل 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 به‌گونه‌ای تنظیم می‌شود که خروجی‌های مطلوب‌تری تولید کند. هم‌ترازسازی رفتار مدل‌های زبانی با ارزش‌ها و انتظارات انسانی، به‌ویژه در کاربردهایی که حساسیت اخلاقی یا عملی دارند، اهمیت فزاینده‌ای یافته است [2]. 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 [2]. 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 به‌دنبال رویکردهایی ساده‌تر، پایدارتر و کم‌هزینه‌تر برای یادگیری ترجیحات انسانی باشند. در همین راستا، الگوریتم Direct Preference Optimization (DPO) معرفی شده است که با حذف کامل مرحله یادگیری تقویتی و مدل پاداش، فرایند هم‌ترازی را به‌شکل مستقیم و مؤثر انجام می‌دهد [1].برای روشن‌تر شدن تفاوت این دو رویکرد، در ادامه ساختار کلی RLHF و DPO به‌صورت شماتیک نمایش داده شده است.



در شکل ۱ ساختار کلی فرایند RLHF نشان داده شده است. در این رویکرد، ترجیحات انسانی به‌عنوان ورودی به یک مدل پاداش ارائه می‌شوند و سپس با استفاده از الگوریتم‌های یادگیری تقویتی، سیاست مدل زبانی به‌صورت پیوسته به‌روزرسانی می‌شود. این چرخهٔ پاداش–سیاست اگرچه قدرت بالایی در یادگیری ترجیحات انسانی دارد، اما به دلیل وجود مدل پاداش جداگانه، نیاز به نمونه‌گیری مکرر از مدل، و به‌کارگیری الگوریتم‌های RL مانند PPO، از نظر محاسباتی بسیار پرهزینه و

گاه ناپایدار است [1]. در مقابل، شکل ۲ رویکرد Direct Preference Optimization (DPO) را نمایش می‌دهد که یک چارچوب ساده‌تر و کارآمدتر برای هم‌ترازی مدل با ترجیحات انسانی ارائه می‌دهد. در DPO مرحلهٔ 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 [1]. در این روش، مدل تنها می‌آموزد احتمال پاسخ ترجیح‌داده‌شده را نسبت به پاسخ مردود افزایش دهد؛ بنابراین یادگیری ترجیحات انسانی بدون نیاز به حلقهٔ Actor–Critic یا مدل پاداش انجام می‌شود. نکته قابل‌توجه این است که DPO همچنان همان هدف بنیادی RLHF—یعنی حداکثرسازی پاداش ضمنی تحت محدودیت واگرایی KL—را دنبال می‌کند، اما این کار را از طریق یک بازنویسی هوشمندانهٔ تابع هدف انجام می‌دهد [1]. 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. 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 [1]. 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—including DPO، ORPO و IPO—از زیان‌های رتبه‌بندی جفتی استفاده می‌کنند که تنها ترتیب نسبی میان پاسخ‌های منتخب و ردشده را حفظ می‌کنند [2]. 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 [2]. 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 که با کالیبره‌کردن پاداش ضمنی نسبت به پاداش پایه، روند یادگیری را پایدارتر کرده و از کاهش ناخواسته احتمال پاسخ منتخب جلوگیری می‌کند [2]. 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 توسعهٔ مدل‌های زبانی بزرگ، مسئلهٔ 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 [2]. بنابراین لازم است فرایند هم‌ترازی نه‌تنها 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 پایهٔ چارچوب Safe-RLHF بنا شده‌اند. در این رویکرد، ابتدا داده‌هایی شامل برچسب‌های «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 [2]. اگرچه Safe-RLHF قادر است رفتارهای نامطلوب را کنترل کند، اما به دلیل آموزش هم‌زمان مدل پاداش، مدل هزینه و حلقهٔ 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 هزینهٔ جداگانه و بدون بهره‌گیری از یادگیری تقویتی محقق کند [2]. در Safe-DPO، داده‌های ترجیحی با استفاده از شاخص‌های ایمنی بازمرتب‌سازی شده و سپس همان فرایند سادهٔ 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 بازیگر–منتقد یا حلقهٔ نمونه‌برداری حذف می‌شود. در ادامه، تفاوت میان Safe-RLHF و Safe-DPO در قالب یک نمودار شماتیک نمایش داده شده است.



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 هزینهٔ جداگانه، سیاست مدل را بر اساس بیشینه‌سازی درست‌نمایی به‌روزرسانی می‌کند که اجزای آبی‌رنگ در شکل نمایانگر آن هستند. در ادامهٔ توسعه‌های انجام‌شده بر روی DPO، الگوریتم Safe-DPO با هدف بهبود ایمنی و پایداری مدل‌های زبانی معرفی شد [1]. 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 [1]. Safe-DPO این محدودیت را برطرف می‌کند و فرآیند 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. نسخهٔ پایهٔ 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 [1]. 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 [1]. 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]. در پاسخ به این نیاز، الگوریتم Unified-PO (Unified Preference Optimization) معرفی شد. 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 [2]. علی‌رغم پیشرفت‌های قابل‌توجه در روش‌های مبتنی بر ترجیحات انسانی، از جمله DPO، Cal-DPO، Safe-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 KL 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 LLMها 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. 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:

شماره	سال	عنوان	منبع انتشار	چکیده کوتاه	نتایج عددی کلیدی	الگوریتم‌های استفاده‌شده	مدل یا چارچوب استفاده‌شده
۱	۲۰۲۳	Direct Preference Optimization: Your Language Model is Secretly a Reward Model	NeurIPS 2023	معرفی DPO به‌عنوان جایگزین ساده‌تر RLHF بدون مدل پاداش	بهبود ۱۰–۱۲٪ در win rate نسبت به PPO در Summarization و Dialogue	DPO (Cross-Entropy)	GPT-2-Large, GPT-J, Pythia
۲	۲۰۲۴	Cal-DPO: Calibrated Direct Preference Optimization	NeurIPS 2024	تنظیم کالیبره برای کاهش نوسان در داده‌های ترجیحی	کاهش ۲۵٪ KL-divergence و بهبود ۸٪ پایداری نسبت به DPO	Calibrated DPO	GPT-J / Anthropic-HH
۳	۲۰۲۵	SGDPO: Self-Guided Direct Preference Optimization	ACL 2025	یادگیری خودراهبری برای کاهش وابستگی به داده انسانی	افزایش ۹٪ در win rate و ۱۲٪ در پایداری نسبت به DPO پایه	Self-Guided DPO	Anthropic-HH, TL;DR
۴	۲۰۲۵	DiffPO: Diffusion-Styled Preference Optimization	ACL 2025	استفاده از ساختار diffusion برای هم‌ترازی LLM ها	کاهش ۲۰٪ زمان استنتاج و حفظ کیفیت در BLEU ≈ 0.87	Diffusion DPO	TL;DR Summarization

۵	۲۰۲۵	Understanding Reference Policies in DPO	NAACL 2025	تحلیل ریاضی تأثیر π_ref بر یادگیری DPO	کاهش ۵٪ loss در policy و KL alignment ≈ 0.31	Analytical DPO	Anthropic-HH
۶	۲۰۲۵	Auto-Weighted GRPO	EMNLP Industry 2025	وزن‌دهی خودکار در multi-objective alignment	بهبود ۱۳٪ در safety score و ۹٪ در helpfulness	GRPO (Auto-Weighted)	CNN/DailyMail Dialogue
۷	۲۰۲۵	Unified Preference Optimization	TMLR 2025	چارچوب یکپارچه برای ترجیحات چنددامنه‌ای	افزایش ۸٪ دقت در cross-domain tasks نسبت به ORPO	Unified DPO	Multi-domain LLMs
۸	۲۰۲۵	Safe-DPO: Enhanced Safety in Preference Optimization	arXiv 2025	نسخهٔ ایمن تر برای داده های نویزدار	کاهش ۲۳٪ خطای پاداش در feedback های نویزدار	Safe-DPO	Reddit + HH
۹	۲۰۲۵	Pre-DPO: Improving Data Utilization	arXiv 2025	استفاده بهینه از داده با مدل مرجع راهنما	افزایش ۱۵٪ کارایی داده و کاهش ۵٪ loss	Pre-DPO	Anthropic-HH
۱۰	۲۰۲۵	Mix- and MoE-DPO: A Variational Inference Approach	arXiv 2025	استفاده از Mixture-of-Experts برای DPO	افزایش ۷٪ دقت و کاهش ۱۰٪ overfitting	MoE-DPO	Anthropic-HH
۱۱	۲۰۲۵	D-RPO: Distributionally Robust Alignment of LLMs	arXiv 2025	مقاوم‌سازی DPO برای داده‌های نامتوازن	افزایش ۱۲٪ robustness score در test set	D-RPO (Robust DPO)	Anthropic-HH

از زمان پیشرفت چشمگیر مدل‌های زبانی بزرگ (LLMs)، مسئله هم‌ترازسازی رفتار این مدل‌ها با ارزش‌ها، استانداردها و ترجیحات انسانی به یکی از محوری‌ترین چالش‌های پژوهش در حوزه هوش مصنوعی تبدیل شده است. روش یادگیری تقویتی از بازخورد انسانی (RLHF) در ابتدا توانست مسیر مناسبی برای این هم‌ترازی ایجاد کند، اما به دلیل نیاز به مدل پاداش، هزینه محاسباتی بسیار بالا، وابستگی به حلقه‌های تقویتی و دشواری جمع‌آوری داده‌های انسانی، به‌سرعت مشخص شد که RLHF به‌تنهایی نمی‌تواند پاسخگوی نیازهای مدل‌های عظیم امروزی باشد. در همین راستا، Rafailov و همکاران الگوریتم بهینه‌سازی ترجیحات مستقیم (DPO) را معرفی کردند [1]؛ روشی که فرآیند یادگیری ترجیح‌محور را بدون نیاز به مدل پاداش، و تنها با استفاده از بازنویسی مسئله مبتنی بر سیاست، به یک تابع زیان ساده و قابل‌محاسبه تبدیل کرد. این

دستاورد، نقطه‌ی عطفی در هم‌ترازی مدل‌های زبانی بود و باعث شد موج قابل‌توجهی از پژوهش‌ها برای توسعه، اصلاح و گسترش این روش شکل بگیرد.

پس از معرفی DPO، یکی از نخستین چالش‌های شناسایی‌شده مسأله کالیبراسیون پاداش‌ها و ناپایداری پارامترهای داخلی بود. پژوهش Cal-DPO نشان داد که مقیاس پاداش‌های ضمنی در DPO با مقیاس پاداش پایه همخوانی ندارد و همین موضوع می‌تواند موجب کاهش احتمال پاسخ‌های منتخب و افت کیفیت مدل شود. Cal-DPO با اعمال اصلاحاتی ساده اما مؤثر، توانست این ضعف را تا حد زیادی برطرف کند و پایداری قابل‌توجهی در نتایج ایجاد نماید [2]. در ادامه، مقالاتی نظیر ORPO و KTO نیز تلاش کردند فرآیند تصمیم‌گیری ترجیح‌محور را از منظر نظری بازتعریف کنند. ORPO به حذف مدل مرجع و کاهش وابستگی به سیاست پایه پرداخت [3] و KTO ترجیحات انسانی را با نظریه چشم‌انداز (Prospect Theory) ادغام کرد و نشان داد که گاهی باید مدل را نسبت به ریسک و عدم قطعیت حساس‌تر آموزش داد [4].

افزون بر این، بخش دیگری از پژوهش‌ها بر بهبود عملکرد DPO در شرایط کم‌نمونه یا ترجیحات پیچیده متمرکز شد. مفهوم ترجیحات گروهی درGRPO [5] چارچوب جدیدی ایجاد کرد که ترجیحات را در قالب خوشه‌های گروهی ساختار می‌دهد و امکان یادگیری چندهدفه را فراهم می‌کند. Soft Preference Optimization [6] نیز با نرم‌سازی توزیع ترجیحات، از سقوط مدل در زمان مواجهه با ترجیحات متناقض جلوگیری کرد و رفتار نرم‌تری در گرادیان‌ها ایجاد کرد. در همین راستا، Pre-DPO [7] رویکردی هدایت‌شده مبتنی بر مدل مرجع ارائه داد که به‌طور خاص برای بهبود کارایی یادگیری در داده‌های محدود طراحی شده بود.

در حوزه‌ی کاربردهای تخصصی‌تر، هم‌ترازی ترجیحی به مدل‌های چندوجهی و مدل‌های کدنویسی نیز گسترش یافت. مقاله PLUM [8] نشان داد که با ترکیب ترجیحات انسانی و ارزیابی اجرای واقعی کد می‌توان مدل‌های کدنویسی را بسیار دقیق‌تر تنظیم کرد. MDPO [9] نیز مسیر تطبیق DPO را به دنیای مدل‌های چندوجهی گسترش داد و نشان داد که ترجیحات می‌توانند برای وظایف تصویر–متن نیز به‌کار گرفته شوند.

از سال ۲۰۲۵ به بعد، جهت‌گیری پژوهش‌ها به‌صورت قابل‌توجهی به سمت پایداری، ایمنی و مقاومت نسبت به داده‌های نویزدار حرکت کرد. مقاله DiffPO [10] با الهام از ساختارهای Diffusion، سرعت و ثبات هم‌ترازی را در مرحله inference افزایش داد. SGDPO [11] نیز با معرفی مفهوم یادگیری خودراهبری (Self-Guided)، نشان داد که بخشی از ترجیحات را می‌توان بدون برچسب‌گذاری انسانی و تنها بر اساس رفتار مدل تولید کرد؛ موضوعی که هزینه انسانی و داده‌ای را به‌شدت کاهش می‌دهد.

در زمینه ایمنی، Safe-DPO [12] با بازمرتب‌سازی ترجیحات بر اساس شاخص‌های ایمنی، DPO را به ابزاری برای تولید پاسخ‌های سالم‌تر و قابل‌اعتمادتر تبدیل کرد، بدون آنکه به مدل پاداش یا ساختارهای پیچیده تقویتی نیاز داشته باشد. در تکمیل این مسیر، Smaug [13] شکست‌های رایج در یادگیری ترجیحی مانند collapse و over-penalization را شناسایی و اصلاح کرد. الگوریتم D-RPO [14] نیز ساختار مقاوم‌سازی‌شده‌ای ارائه داد که عملکرد DPO را در برابر توزیع‌های نامتوازن، داده‌های نویزدار و ترجیحات ناسازگار بهبود می‌بخشد.

افزون بر این، پژوهش‌های مهمی نیز در راستای گسترش چارچوب نظری DPO انجام شده است. Unified Preference Optimization (UPO) [15] نشان داد که بسیاری از الگوریتم‌های ترجیح‌محور در واقع نسخه‌هایی از یک چارچوب یکپارچه هستند و می‌توان آن‌ها را تحت یک تابع هدف عمومی فرمول‌بندی کرد. این موضوع مسیر توسعه و مقایسه مدل‌ها را ساده‌تر و منسجم‌تر می‌کند. Towards Efficient Exact Optimization [16] نیز تلاش کرده است که هم‌ترازی ترجیح‌محور را از نظر زمانی و حافظه‌ای کارآمدتر کند. در همین راستا، Mix- and MoE-DPO [17] با آوردن معماری Mixture-of-Experts به فضای ترجیحی، امکان مدل‌سازی ترجیحات پیچیده و چندبُعدی را فراهم کرده است. مقاله Self-Rewarding Language Models [18] نیز نشان داد که مدل‌ها می‌توانند بخشی از پاداش آموزشی موردنیاز خود را به‌صورت ضمنی تولید کنند، موضوعی که زمینه را برای روش‌های خودتنظیمی و کاهش نیاز به بازخورد انسانی فراهم می‌کند.

با وجود تمام این پیشرفت‌ها، یک مسئله‌ی بنیادین در میان همه پژوهش‌های بررسی‌شده مشترک است: وابستگی تمامی نسخه‌های موجود به ضریب ثابت β. β پارامتر کلیدی‌ای است که تعادل میان «افزایش احتمال پاسخ ترجیح‌داده‌شده» و «حفظ فاصلهٔ منطقی از سیاست مرجع» را کنترل می‌کند. ثابت ماندن β موجب ناپایداری گرادیان، نوسانات شدید در فرآیند یادگیری، حساسیت بالا نسبت به داده‌های نویزدار و کاهش کیفیت خروجی می‌شود [2, 11-13]. در هیچ‌یک از ۲۱ مقاله مرور شده، یک راهکار رسمی و عملی برای تنظیم تطبیقی β ارائه نشده است.

به‌همین دلیل، پژوهش حاضر تلاش می‌کند نسخه‌ای جدید از DPO تحت عنوان Adaptive-DPO طراحی کند که در آن β به‌صورت پویا، هوشمند و متناسب با وضعیت لحظه‌ای مدل تنظیم می‌شود. انتظار می‌رود این روش پایداری آموزشی، دقت در تولید پاسخ و کیفیت خروجی مدل‌های زبانی—به‌ویژه در حوزهٔ دستیارهای هوشمند کدنویسی—را بهبود دهد. این پژوهش در امتداد منطقی مسیر توسعه DPO قرار دارد و یکی از مهم‌ترین خلأهای علمی موجود را هدف قرار می‌دهد.

روش پیشنهادی:

در این پژوهش، روشی جدید برای بهبود پایداری و کارایی الگوریتم Direct Preference Optimization (DPO) 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 می‌ماند. با این حال، β ثابت در عمل منجر به مشکلاتی از جمله حساسیت بالا به مقدار انتخابی، نوسان در همگرایی، و افزایش ناخواسته‌ی KL-divergence نسبت به سیاست مرجع می‌شود. این مسئله می‌تواند باعث drift از مدل مرجع، افت کیفیت پاسخ‌ها و گاهی بروز پدیده‌ی model collapse شود. برای رفع این چالش‌ها، در این تحقیق الگوریتمی با عنوان Adaptive-β DPO معرفی می‌شود که در آن مقدار β در هر مرحله بر اساس رفتار مدل، میزان اختلاف بین پاسخ‌های ترجیحی و غیرترجیحی، و فاصله‌ی مدل از سیاست مرجع، به‌صورت پویا تنظیم می‌گردد.

ایده‌ی بنیادین این پژوهش بر این فرض بنا شده است که شدت اعمال ترجیحات انسانی نباید در طول آموزش ثابت بماند، زیرا مدل در مراحل مختلف یادگیری رفتارهای متفاوتی نشان می‌دهد. زمانی که مدل بیش‌ازحد از سیاست مرجع دور شده باشد، لازم است قدرت تغییرات (β) کاهش یابد تا از افزایش KL و بی‌ثباتی جلوگیری شود. در مقابل، هنگامی که مدل هنوز تفاوت کافی بین پاسخ‌های ترجیحی و غیرترجیحی ایجاد نکرده باشد، باید β افزایش یابد تا یادگیری ترجیحات مؤثرتر صورت گیرد. بنابراین، β باید متناسب با شرایط لحظه‌ای مدل تنظیم شود.

در روش پیشنهادی، پس از دریافت هر ورودی شامل یک پرامپت و دو پاسخ (پاسخ ترجیحی و ناترجیحی)، مدل احتمال شرطی تولید هر پاسخ را محاسبه می‌کند. اختلاف این دو مقدار به عنوان یک شاخص «اطمینان مدل» نسبت به ترجیح انسانی در آن نمونه در نظر گرفته می‌شود. سپس KL-divergence بین سیاست فعلی مدل و سیاست مرجع محاسبه شده و میزان دور شدن مدل از رفتار اولیه سنجیده می‌شود. ضریب β با توجه به این دو مقدار و همچنین مرحله جاری آموزش به‌صورت تطبیقی بازتنظیم می‌شود. این مقدار جدید β وارد فرمول DPO شده و در محاسبه تابع زیان مورد استفاده قرار می‌گیرد. به این ترتیب، مدل در نمونه‌هایی که سخت‌تر هستند یا شک بیشتری نسبت به ترجیح صحیح دارد، به‌صورت محافظه‌کارانه‌تری به‌روزرسانی می‌شود و در نمونه‌های ساده‌تر یا زمانی که مدل نسبت به ترجیح انسانی اطمینان بیشتری دارد، یادگیری سریع‌تر صورت می‌گیرد.

برای تعیین مقدار پویا، ضریب β بر اساس سه مؤلفه محاسبه می‌شود. اول اختلاف لگاریتمی بین پاسخ ترجیحی و ناترجیحی. هرچه این اختلاف بیشتر باشد، مدل برای تغییرات شدیدتر آمادگی بیشتری دارد. دوم KL-divergence نسبت به سیاست مرجع. افزایش KL نشان‌دهنده drift است و باید منجر به کاهش β شود. و سوم مرحله‌ی فعلی آموزش در مراحل ابتدایی β ملایم‌تر انتخاب می‌شود و به‌تدریج افزایش می‌یابد.با استفاده از این عوامل، فرمولی تطبیقی برای β به‌صورت زیر پیشنهاد می‌شود:

$$\beta_1 = \frac{\alpha \,.\, gap_t}{1 + \mathrm{KL}_t}$$

که در آن α یک ضریب تنظیمی کوچک است. این معادله امکان کنترل پویا و معنادار میزان تأثیرگذاری ترجیحات انسانی را فراهم می‌کند و تعادل بهتری بین یادگیری و پایداری ایجاد می‌نماید.

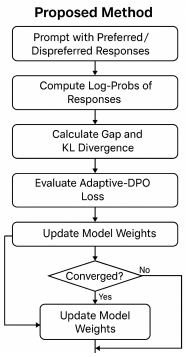
پس از تعیین β تطبیقی، تابع زیان اصلی DPO به‌صورت زیر بازنویسی می‌شود:

$$L_{Adoptive-DPO} = -\mathrm{E}[\log \sigma\,(\beta_t \,.\, gap_t)]$$

این تابع زیان باعث می‌شود که در هر مرحله از آموزش شدت ترجیحات انسانی متناسب با وضعیت مدل تنظیم شود، بدون آنکه نیاز به طراحی الگوریتم‌های پیچیده‌ی تقویتی وجود داشته باشد.

به طور خلاصه این روش از دریافت ورودی و پاسخ های ترجیحی / ناترجیحی شروع و در ادامه محاسبه log-prob برای هر پاسخ و محاسبه اختلاف این مقادیر به عنوان gap، اندازه گیری KL-divergence نسبت به مدل مرجع، تعیین مقدار β جدید، بروزرسانی وزن های مدل و در آخر تکرار فرایند تا همگرایی.

روش Adaptive-β DPO دارای مزایا های افزایش پایداری آموزش به‌ویژه در مدل‌های بزرگ، جلوگیری از drift و کنترل دقیق KL، تقویت یادگیری ترجیحات سخت و عدم overshoot در نمونه های حساس، سازگاری با داده های متنوع و توزیع های مختلف و امکان استفاده بدون نیاز به روش های پیچیده تقویتی.



[3-11, 13, 14, 16-31]

منابع

1. 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.
2. 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.
3. Hong, J., N. Lee, and J. Thorne, *Orpo: Monolithic preference optimization without reference model*. arXiv preprint arXiv:2403.07691, 2024.
4. Ethayarajh, K., et al., *Kto: Model alignment as prospect theoretic optimization, 2024*. URL https://arxiv. org/abs/2402.01306.

5. Zhao, S., J. Dang, and A. Grover, *Group preference optimization: Few-shot alignment of large language models*. arXiv preprint arXiv:2310.11523, 2023.
6. Sharifnassab, A., et al., *Soft preference optimization: Aligning language models to expert distributions*. arXiv preprint arXiv:2405.00747, 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. 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.
9. Wang, F., et al., *mdpo: Conditional preference optimization for multimodal large language models*. arXiv preprint arXiv:2406.11839, 2024.
10. 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.
11. Zhu, W., et al., *SGDPO: Self-Guided Direct Preference Optimization for Language Model Alignment*. arXiv preprint arXiv:2505.12435, 2025.
12. Kim, G.-H., et al., *SafeDPO: A simple approach to direct preference optimization with enhanced safety*. arXiv preprint arXiv:2505.20065, 2025.
13. Pal, A., et al., *Smaug: Fixing failure modes of preference optimisation with dpo-positive, 2024*. URL https://arxiv. org/abs/2402.13228.
14. Wu, J., et al., *Towards robust alignment of language models: Distributionally robustifying direct preference optimization*. arXiv preprint arXiv:2407.07880, 2024.
15. Badrinath, A., P. Agarwal, and J. Xu, *Unified Preference Optimization: Language Model Alignment Beyond the Preference Frontier*. arXiv preprint arXiv:2405.17956, 2024.
16. Ji, H., *Towards efficient exact optimization of language model alignment (2024)*. URL https://arxiv. org/abs/2402.00856. **2402**.
17. Bohne, J., et al., *Mix-and MoE-DPO: A Variational Inference Approach to Direct Preference Optimization*. arXiv preprint arXiv:2510.08256, 2025.
18. Yuan, W., et al. *Self-rewarding language models*. in *Forty-first International Conference on Machine Learning*. 2024.
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. Liu, Y., P. Liu, and A. Cohan, *Understanding reference policies in direct preference optimization*. arXiv preprint arXiv:2407.13709, 2024.
22. Xiao, W., et al., *A comprehensive survey of direct preference optimization: Datasets, theories, variants, and applications*. arXiv preprint arXiv:2410.15595, 2024.
23. 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.

24. Liang, X., et al., *ROPO: Robust Preference Optimization for Large Language Models*. arXiv preprint arXiv:2404.04102, 2024.
25. Liu, S., et al., *A survey of direct preference optimization*. arXiv preprint arXiv:2503.11701, 2025.
26. He, J., H. Yuan, and Q. Gu, *Accelerated preference optimization for large language model alignment*. arXiv preprint arXiv:2410.06293, 2024.
27. Zeng, D., et al. *On diversified preferences of large language model alignment*. in *Findings of the association for computational linguistics: EMNLP 2024*. 2024.
28. Sun, S., et al., *Reward-aware preference optimization: A unified mathematical framework for model alignment*. arXiv preprint arXiv:2502.00203, 2025.
29. Lu, J., et al., *Adavip: Aligning multi-modal llms via adaptive vision-enhanced preference optimization*. arXiv preprint arXiv:2504.15619, 2025.
30. 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.
31. Long, O., et al., *Training language models to follow instructions with human feedback*. Advances in neural information processing systems, 2022. **35**: p. 27730–27744.