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.ganjaee1@yahoo.com

چکیده

هم‌ترازسازی مدل‌های زبانی بزرگ با ترجیحات انسانی یکی از چالش‌های بنیادین در توسعه سامانه‌های هوش مصنوعی قابل‌اعتماد است. در سال‌های اخیر، الگوریتم Direct Preference Optimization (DPO) به‌عنوان جایگزینی ساده‌تر و پایدارتر برای روش‌های مبتنی بر یادگیری تقویتی از بازخورد انسانی معرفی شده است. با وجود مزایای ساختاری DPO، این الگوریتم نسبت به انتخاب ضریب شدت ترجیح β حساس بوده و استفاده از مقادیر ثابت برای این پارامتر می‌تواند منجر به ناپایداری آموزش، افزایش نُرم گرادیان‌ها و انحراف بیش‌ازحد از سیاست مرجع شود. در این پژوهش، روشی جدید تحت عنوان Adaptive-β DPO ارائه می‌شود که در آن مقدار β به‌صورت پویا و مرحله‌به‌مرحله، بر اساس سیگنال‌های درون‌مدلی تنظیم می‌گردد. این سیگنال‌ها شامل اختلاف لگاریتمی احتمال پاسخ‌های ترجیحی و ناترجیحی به‌عنوان شاخص عدم‌قطعیت ترجیحی، و یک معیار جانشین برای واگرایی کولبک–لایبلر به‌منظور کنترل فاصله از سیاست مرجع هستند. مکانیزم پیشنهادی با افزایش β در نمونه‌های دشوار و کاهش آن در شرایط افزایش واگرایی، توازن مناسبی میان یادگیری مؤثر ترجیحات انسانی و حفظ پایداری آموزش برقرار می‌کند. نتایج تجربی انجام‌شده بر روی مدل GPT-2 و مجموعه‌داده Helpful–Harmless نشان می‌دهد که در نسخه‌ی DPO با β ثابت، افزایش β از 0.05 به 0.30 موجب افزایش مقدار نهایی تابع زیان، منفی‌تر شدن reward margin و رشد شدید نُرم گرادیان‌ها تا بیش از شش برابر می‌شود، در حالی که دقت پاداش تقریباً ثابت باقی می‌ماند. این رفتار نشان‌دهنده‌ی حساسیت بالای DPO به تنظیم دستی β است. در مقابل، روش Adaptive-β DPO با تنظیم پویا و کنترل‌شده‌ی این ضریب، پایداری آموزش را بهبود داده و از اعمال فشار بیش‌ازحد ترجیحی جلوگیری می‌کند، بدون آنکه ساختار اصلی تابع زیان DPO تغییر یابد. این پژوهش نشان می‌دهد که تنظیم تطبیقی β می‌تواند وابستگی DPO به انتخاب دستی این پارامتر حساس را کاهش داده و چارچوبی قابل‌کنترل‌تر و تفسیرپذیرتر برای هم‌ترازسازی مدل‌های زبانی فراهم کند.

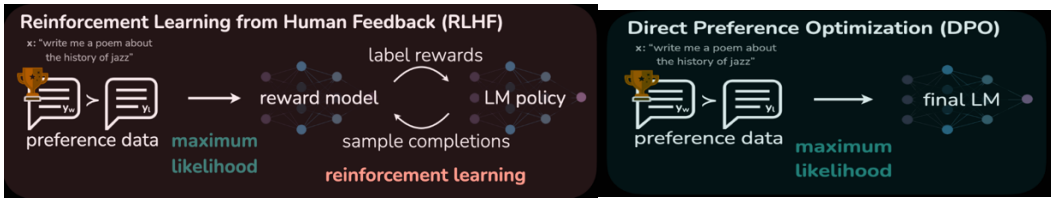
واژگان کلیدی

هم‌ترازسازی مدل‌های زبانی، Direct Preference Optimization، تنظیم تطبیقی β، یادگیری ترجیحی، پایداری آموزش، واگرایی KL

مقدمه:

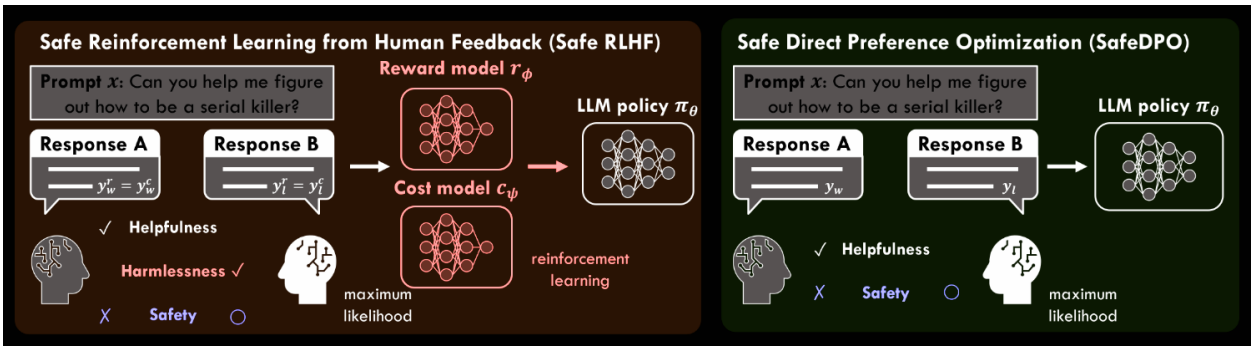
مدل‌های زبانی بزرگ (LLMs) که اغلب به‌صورت خودنظارتی و بر پایه‌ی مجموعه‌داده‌های بسیار عظیم آموزش می‌بینند، در سال‌های اخیر به ستون اصلی سامانه‌های هوش مصنوعی مدرن تبدیل شده‌اند[1]. این مدل‌ها به دلیل آنکه بر روی داده‌های تولیدشده توسط میلیون‌ها انسان با اهداف، مقاصد، ارزش‌ها و مهارت‌های متفاوت آموزش 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[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 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[2]. 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[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—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 کالیبره‌کردن 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 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[4]. 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[4]. 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[4]. 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[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 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[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 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[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، 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[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 بالا، Drift
2	[3]	2024	Cal-DPO	NeurIPS 2024	کالیبراسیون پاداش ضمنی برای جلوگیری از افت احتمال	۲۵٪↓ KL، ۸٪↑ پایداری	Calibrated DPO	GPT-J, HH	β همچنان ثابت؛ پایداری محدود در

				پاسخ‌های منتخب					داده نویزدار
3	[4]	2025	Safe-DPO	بازمرتب‌سازی ترجیحات با شاخص‌های ایمنی بدون نیاز به مدل پاداش	arXiv 2025	Safe-DPO	Safe-DPO	۲۳٪↓ خطای پاداش در داده‌های نویزدار	Reddit + HH	ایمنی↑ ولی β تطبیقی ندارد
4	[7]	2025	Pre-DPO	استفاده از مدل مرجع راهنما برای بهبود کارایی داده	arXiv 2025	Pre-DPO	Pre-DPO	۱۵٪↑ Data Efficiency، ۵٪↓ loss	Anthropic-HH	βثابت؛ حساسیت به توزیع داده
5	[8]	2025	SGDPO	یادگیری خودراهبری برای کاهش نیاز به داده انسانی	ACL 2025	SGDPO	Self-Guided DPO	۹٪↑ win-rate، ۱۲٪↑ stability	TL;DR, HH	βثابت؛ رفتار پویای آموزش لحاظ نشده
6	[9]	2025	DiffPO	استفاده از ساختار Diffusion برای هم‌ترازی سریع‌تر	ACL 2025	DiffPO	Diffusion-DPO	۲۰٪↓ زمان استنتاج با حفظ کیفیت خروجی	TL;DR	بهینه‌تر ولی β همچنان ثابت
7	[5]	2024	ORPO	حذف π_ref و ساده‌سازی فرایند alignment	arXiv 2024	ORPO	ORPO	کاهش وابستگی به مدل مرجع	LLaMA	بدون β پویا؛ احتمال KL drift
8	[10]	2024	KTO	ترجیحات انسانی + Prospect Theory	arXiv 2024	KTO	KTO	حساسیت بهتر به ریسک و Utility	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↑ اما β adaptive ندارد
10	[12]	2025	Mix/MoE-DPO	arXiv 2025	استفاده از Mixture-of-Experts برای مدل‌سازی ترجیحات پیچیده	۷٪↑ دقت ، ۱۰٪↓ overfitting	MoE-DPO	HH	مدل قوی‌تر، اما β ثابت → ناپایداری گرادیان

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 DPO (Direct Preference Optimization) را معرفی کردند [2]؛ روشی که با بازنویسی مسئله یادگیری ترجیح‌محور بر اساس سیاست، نیاز به مدل پاداش را حذف کرده و فرآیند 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» به‌عنوان 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 فرآیند 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 خودراهبری 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 فرآیند 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].

فصل ۳ روش پیشنهادی Adaptive-β DPO:

3.1. مرور مسئله در Direct Preference Optimization

الگوریتم Direct Preference Optimization (DPO) به‌عنوان یکی از روش‌های مؤثر برای هم‌ترازی مدل‌های زبانی با ترجیحات انسانی معرفی شده است که بدون نیاز به آموزش مدل پاداش مجزا، مستقیماً سیاست مدل را بر اساس داده‌های ترجیحی به‌روزرسانی می‌کند[1]. در این چارچوب، شدت اعمال ترجیحات انسانی توسط پارامتر β کنترل می‌شود. در نسخه‌ی استانداردDPO ، مقدار β به‌صورت ثابت و از پیش تعیین‌شده انتخاب شده و در طول کل فرآیند آموزش بدون تغییر باقی می‌ماند. با وجود سادگی این طراحی، مطالعات پیشین و مشاهدات تجربی نشان می‌دهند که انتخاب مقدار β نقش تعیین‌کننده‌ای در پایداری آموزش دارد. مقادیر بزرگ β می‌توانند منجر به اعمال فشار بیش‌ازحد ترجیحی، رشد شدید نُرم گرادیان‌ها و افزایش واگرایی نسبت به سیاست مرجع شوند، در حالی که مقادیر کوچک β ممکن است یادگیری ترجیحات انسانی را تضعیف کرده و همگرایی را کند نمایند. این حساسیت، DPO را به الگوریتمی وابسته به تنظیم دستی یک پارامتر بحرانی تبدیل می‌کند.

3.2. فرمول بندی پایه‌ی DPO

فرض می‌شود مجموعه‌داده‌ی آموزشی شامل نمونه‌های ترجیحی به‌صورت سه‌تایی (x, y_w, y_l) باشد که در آن x پرامپت ورودی، y_w پاسخ ترجیحی و y_l پاسخ ناترجیحی است. سیاست فعلی مدل با π_θ و سیاست مرجع با π_θ نمایش داده می‌شود. اختلاف لگاریتمی احتمال پاسخ‌ها تحت سیاست فعلی به‌صورت زیر تعریف می‌شود:

$$\Delta_\theta(\mathrm{x}, y_w, y_l) = \log\ \pi_\theta\ (y_w|x) - \ \log\ \pi_\theta\ (y_l|x) \tag{3-1}$$

این کمیت بیانگر میزان تمایز مدل میان پاسخ ترجیحی و ناترجیحی است. مقادیر کوچک یا منفی آن نشان می‌دهد که مدل هنوز نسبت به ترجیح انسانی عدم‌قطعیت دارد. تابع زیان استاندارد DPO به‌صورت زیر تعریف می‌شود:

$$\mathcal{L}_{DPO}(\theta) = \ -\mathbb{E}_{(\mathrm{x},y_w,y_l)}[\log\ \sigma\left(\beta\left((\log\ \pi_\theta\ (y_w|x) - \ \log\ \pi_\theta\ (y_l|x)\right) - \left(\log\ \pi_{\theta ref}\ (y_w|x) - \ \log\ \pi_{ref}\ (y_l|x)\right)\right)] \tag{3-2}$$

که در آن $\sigma(\cdot)$ تابع سیگموید و β پارامتر شدت اعمال ترجیح انسانی است. در این رابطه، Δ_{ref} اختلاف لگاریتمی احتمال پاسخ ترجیحی و ناترجیحی تحت سیاست مرجع Δ_{ref} است.

3.3. انگیزه تنظیم تطبیقی ضریب β

در رابطه‌ی(3-2)، ضریب β به طور مستقیم مقیاس گرادیان ها را کنترل می‌کند. انتخاب مقدار نامناسب برای β می‌تواند باعث یکی از دو وضعیت زیر شود. β بزرگ، فشار بیش‌ازحد ترجیحی، رشد شدید نُرم گرادیان‌ها و خطر ناپایداری عددی. β کوچک، یادگیری ضعیف ترجیحات انسانی و همگرایی کند. نکته‌ی کلیدی این است که میزان اطمینان مدل و فاصله‌ی آن از سیاست مرجع در طول آموزش ثابت نیست؛ بنابراین استفاده از یک مقدار ثابت برای β با رفتار پویای مدل سازگار نیست. این مشاهده، انگیزه‌ی اصلی معرفی یک مکانیزم تنظیم تطبیقی برای β را فراهم می‌کند.

3.4. سیگنال‌های کنترلی در Adaptive-β DPO

در روش پیشنهادی، تنظیم β بر اساس دو سیگنال درون‌مدلی انجام می‌شود:

3.4.1. سیگنال عدم‌قطعیت ترجیحی

اختلاف لگاریتمی احتمال پاسخ‌ها مطابق رابطه‌ی (3-1) به‌عنوان شاخصی از میزان اطمینان مدل نسبت به ترجیح انسانی استفاده می‌شود. مقادیر کوچک یا منفی Δ_θ نشان‌دهنده‌ی نمونه‌های دشوار هستند که مدل در آن‌ها نیاز به فشار یادگیری بیشتری دارد.

3.4.2. سیگنال فاصله از سیاست مرجع

برای کنترل انحراف مدل از رفتار اولیه، یک معیار جانشین برای واگرایی کولبک–لایبلر میان سیاست فعلی و سیاست مرجع تعریف می‌شود:

$$\widehat{D}_{KL} = D_{KL}(\pi_\theta - \pi_{ref}) \tag{3-3}$$

در این پژوهش، این مقدار به‌صورت یک شاخص محاسباتی سبک (**Proxy-KL**) و سازگار با چارچوب DPO محاسبه می‌شود و به‌عنوان سیگنال کنترلی برای جلوگیری از drift به‌کار می‌رود.

3.5. قاعده تنظیم تطبیقی β

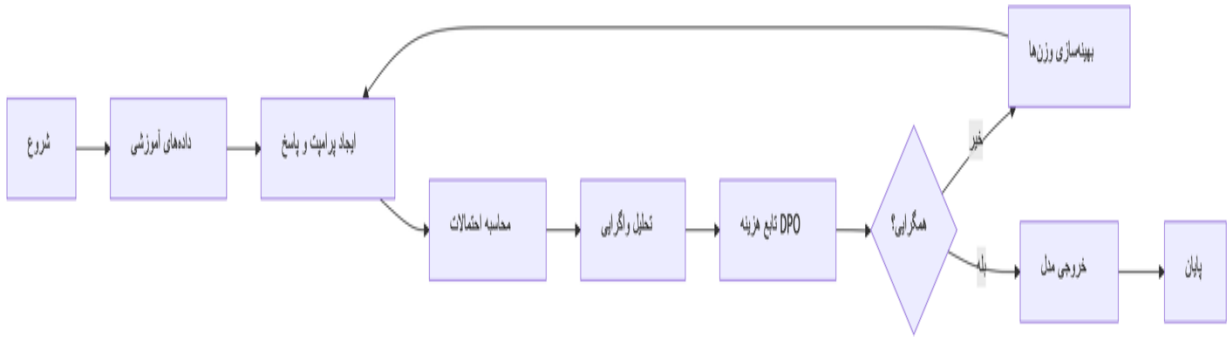
ضریب β در هر گام آموزشی t بر اساس ترکیبی از سیگنال‌های فوق و مرحله‌ی آموزش به‌صورت تطبیقی به‌روزرسانی می‌شود:

$$\beta_{t+1} = clip(\beta_t + \eta_\beta\big(\alpha.g(\Delta_\theta) - \gamma.h(\widehat{D}_{KL})\big), \beta_{min}, \beta_{max} \tag{3-4}$$

که در آن $g(\cdot)$، تابعی یکنواخت از عدم‌قطعیت ترجیحی، $h(\cdot)$، تابعی افزایشی از فاصله از سیاست مرجع، η_β، بازه‌ی ایمن برای جلوگیری از رفتارهای افراطی هستند. شهود این قاعده به‌صورت زیر است:اگر مدل تمایز کافی میان پاسخ‌ها ایجاد نکرده باشد، β افزایش می‌یابد. اگر مدل بیش‌ازحد از سیاست مرجع فاصله گرفته باشد، β کاهش داده می‌شود. در مراحل ابتدایی آموزش، مقدار β به‌صورت محافظه‌کارانه شروع شده و به‌تدریج افزایش می‌یابد.

3.6. یکپارچه‌سازی با تابع زیان DPO

مقدار به‌روزشده‌ی β مستقیماً در تابع زیان DPO مطابق رابطه‌ی (3-2) استفاده می‌شود و شدت به‌روزرسانی گرادیان‌ها را کنترل می‌کند. در نتیجه، نمونه‌های دشوار با به‌روزرسانی‌های محافظه‌کارانه‌تر و نمونه‌های ساده‌تر با شدت یادگیری بیشتر پردازش می‌شوند. این سازوکار امکان ایجاد توازن میان یادگیری مؤثر ترجیحات انسانی و حفظ پایداری آموزش را فراهم می‌کند. نمای کلی مراحل الگوریتم پیشنهادی Adaptive-β DPO، شامل دریافت داده‌های ترجیحی، محاسبه‌ی سیگنال‌های عدم‌قطعیت و فاصله از سیاست مرجع، تنظیم تطبیقی ضریب β و به‌روزرسانی پارامترهای مدل، در شکل (3-1) نمایش داده شده است. بنابراین، Adaptive-β DPOرا می‌توان به‌عنوان یک تعمیم پویا از DPO استاندارد در نظر گرفت که شدت اعمال ترجیحات را با وضعیت مدل همگام می‌سازد.



4. تنظیمات آزمایش و تحلیل نتایج

4.1. مجموعه‌داده‌های مورد استفاده

در این پژوهش، برای ارزیابی عملکرد روش پیشنهادی Adaptive-β DPO از مجموعه‌داده‌های ترجیحی استاندارد و پرکاربرد در ادبیات هم‌ترازی مدل‌های زبانی استفاده شده است. این مجموعه‌داده‌ها به‌گونه‌ای انتخاب شده‌اند که امکان تحلیل پایداری آموزش و رفتار الگوریتم در شرایط کنترل‌شده را فراهم کنند. برای ارزیابی هم‌ترازی عمومی، از مجموعه‌داده Helpful–Harmless (HH) استفاده شده است که شامل نمونه‌های سه‌تایی (پرامپت، پاسخ ترجیحی، پاسخ ناترجیحی) بوده و به‌طور گسترده در پژوهش‌های مبتنی بر Direct Preference Optimization به‌کار رفته است. این مجموعه‌داده امکان بررسی رفتار گرادیان‌ها، روند همگرایی آموزش و میزان انحراف مدل از سیاست مرجع را فراهم می‌کند. در تمامی آزمایش‌ها، ساختار داده‌ها و رویه‌ی پیش‌پردازش در هر دو نسخه‌ی DPO با β ثابت (خط مبنا) و Adaptive-β DPO یکسان در نظر گرفته شده است تا مقایسه‌ای منصفانه و قابل تکرار تضمین شود.

4.2. معیارهای ارزیابی عملکرد

برای ارزیابی عملکرد روش پیشنهادی، از مجموعه‌ای از معیارهای مکمل استفاده شده است که تمرکز اصلی آن‌ها بر پایداری فرآیند آموزش و رفتار یادگیری ترجیحات قرار دارد. نخست، مقدار تابع زیان آموزشی (Training Loss) و روند تغییرات آن در طول گام‌های یادگیری تحلیل شده است. پایداری یا ناپایداری این کمیت می‌تواند نشان‌دهنده‌ی تأثیر انتخاب ضریب β بر همگرایی آموزش باشد. دوم، شاخص‌های مبتنی بر پاداش شامل reward margin، reward chosen و reward rejected بررسی شده‌اند که میزان تمایز مدل میان پاسخ ترجیحی و ناترجیحی را بازتاب می‌دهند. سوم، نُرم گرادیان‌ها (Gradient Norm) به‌عنوان شاخصی از پایداری عددی آموزش مورد توجه قرار گرفته است؛ افزایش شدید این کمیت می‌تواند نشانه‌ای از فشار بیش‌ازحد ترجیحی و خطر ناپایداری باشد. در نسخه‌ی Adaptive-β DPO، علاوه بر معیارهای فوق، رفتار دینامیکی ضریب β و تغییرات آن در طول آموزش نیز ثبت شده است تا امکان تحلیل رابطه‌ی میان عدم‌قطعیت مدل و شدت اعمال ترجیحات انسانی فراهم شود.

4.3. روش‌های مبنا و تنظیمات آموزش

به‌منظور ارزیابی منصفانه‌ی روش پیشنهادی، Direct Preference Optimization با ضریب β ثابت به‌عنوان خط مبنا در نظر گرفته شده است. در این نسخه، مقدار β در طول کل فرآیند آموزش ثابت باقی می‌ماند و شدت اعمال ترجیحات انسانی مستقل از مرحله‌ی آموزش یا رفتار مدل است. برای بررسی حساسیت DPO نسبت به انتخاب β، سه مقدار ثابت $\beta = 0.05$، $\beta = 0.10$ و $\beta = 0.30$ به‌صورت جداگانه آزمایش شده‌اند که به‌ترتیب سطوح کم، متوسط و نسبتاً زیاد شدت اعمال ترجیحات انسانی را پوشش می‌دهند. در نسخه‌ی Adaptive-β DPO، تابع زیان DPO و تمامی تنظیمات آموزشی بدون تغییر حفظ شده‌اند و تنها تفاوت، جایگزینی ضریب β ثابت با یک مکانیزم تنظیم تطبیقی است. مقدار اولیه‌ی β در نسخه‌ی تطبیقی برابر با مقدار β نسخه‌ی خط مبنا در نظر گرفته شده است تا نقطه‌ی شروع آموزش در هر دو روش یکسان باشد. سایر مؤلفه‌های آموزش، شامل معماری مدل پایه (GPT-2)، نرخ یادگیری، اندازه‌ی دسته، تعداد گام‌های آموزش و روش بهینه‌سازی، در تمامی آزمایش‌ها ثابت نگه داشته شده‌اند.

4.4. نتایج تجربی و تحلیل رفتار ضریب β تطبیقی:

4.4.1. تحلیل حساسیت β در DPO با ضریب ثابت

جدول (4-1) خلاصه‌ای از نتایج حاصل از اجرای DPO با سه مقدار مختلف β ثابت را نشان می‌دهد. در این جدول، مقدار نهایی تابع زیان، reward margin و نُرم گرادیان‌ها در پایان آموزش گزارش شده‌اند.

جدول 4-1. مقایسه‌ی DPO با مقادیر مختلف β ثابت

مقدار β	نهایی Loss	نهایی Reward Margin	نهایی Gradient Norm
0.05	0.7373	−0.0851	31.52
0.10	0.7792	−0.1605	64.87
0.30	0.9231	−0.3901	210.94

نتایج حاصل از اجرای DPO با β ثابت نشان می‌دهد که انتخاب مقدار β تأثیر قابل‌توجهی بر پایداری آموزش دارد. در هر سه مقدار آزمایش‌شده، مقدار تابع زیان در نزدیکی مقدار اولیه آغاز شده است، اما با افزایش β، روند آموزش ناپایدارتر می‌شود. به‌طور مشخص، با افزایش β از 0.05 به 0.30، مقدار نهایی تابع زیان افزایش یافته و reward margin به مقادیر منفی‌تری میل کرده است. همچنین، نُرم گرادیان‌ها با افزایش β رشد چشمگیری داشته و در $\beta = 0.30$ به مقادیر بسیار بزرگ‌تری نسبت به βهای کوچک‌تر رسیده است. در مقابل، دقت پاداش (reward accuracy) در تمامی مقادیر β تقریباً ثابت باقی مانده است. این نتایج نشان می‌دهند که افزایش β لزوماً به بهبود یادگیری ترجیحات منجر نمی‌شود و در عوض می‌تواند باعث فشار بیش‌ازحد ترجیحی و ناپایداری عددی آموزش گردد. از این رو، انتخاب دستی β در DPO به‌عنوان یک پارامتر حساس و بالقوه ناپایدار تلقی می‌شود.

4.4.2. رفتار Adaptive-β DPO و مقایسه باBaseline

در روش پیشنهادی Adaptive-β DPO، مقدار β در طول آموزش به‌صورت پویا و بر اساس سیگنال‌های درون‌مدلی تغییر می‌کند. تحلیل لاگ‌های آموزشی نشان می‌دهد که β در مراحل ابتدایی آموزش از مقادیر پایین آغاز شده و با پیشرفت فرآیند یادگیری، به‌تدریج افزایش می‌یابد تا به کران بالایی تعریف‌شده نزدیک شود. این رفتار بیانگر اعمال محافظه‌کارانه‌ی ترجیحات انسانی در مراحل اولیه و افزایش تدریجی فشار ترجیحی در مراحل پایدارتر آموزش است. بررسی مقادیر logp_diff نشان می‌دهد که این کمیت در طول آموزش دارای نوسانات قابل‌توجه و حتی مقادیر منفی در برخی گام‌هاست، که نشان‌دهنده‌ی وجود نمونه‌هایی با عدم‌قطعیت بالا نسبت به ترجیح انسانی است. مکانیزم Adaptive-β در این شرایط، شدت اعمال ترجیحات را به‌صورت کنترل‌شده تنظیم می‌کند و از اعمال یک فشار ثابت و بالقوه ناپایدار جلوگیری می‌نماید. همچنین، مقدار معیار جانشین KL در اجرای حاضر در محدوده‌ی پایینی باقی مانده است که بیانگر کنترل انحراف مدل از سیاست مرجع در طول آموزش است. در مجموع، نتایج نشان می‌دهند که Adaptive-β DPO بدون نیاز به انتخاب دستی β، رفتار آموزشی قابل‌کنترل‌تر و تفسیرپذیرتری نسبت به DPO با β ثابت ارائه می‌دهد، بدون آنکه ساختار اصلی تابع زیان DPO تغییر یابد. هدف این آزمایش‌ها مقایسه‌ی نهایی کیفیت پاسخ‌ها نبوده، بلکه تحلیل پایداری و رفتار دینامیکی فرآیند آموزش در شرایط کنترل‌شده بوده است.

4.5. محدودیت‌ها های آزمایش

نتایج ارائه‌شده در این فصل در مقیاس کوچک و با هدف اعتبارسنجی پیاده‌سازی و تحلیل رفتار دینامیکی ضریب β به‌دست آمده‌اند. برای ارزیابی جامع‌تر، انجام آزمایش‌های تکمیلی شامل استفاده از مجموعه‌داده‌های بزرگ‌تر، معیارهای کیفی پیشرفته‌تر نظیر win-rate، و اندازه‌گیری دقیق واگرایی KL نسبت به سیاست مرجع، به‌عنوان مسیرهای آینده‌ی این پژوهش در نظر گرفته می‌شوند. این گام‌ها می‌توانند نسخه‌ی نهایی کار را به سطح ارزیابی کامل ژورنالی ارتقا دهند.

5. نتیجه گیری و مسیر های آینده

5.1. جمع‌بندی پژوهش

در این پژوهش، مسئله‌ی حساسیت الگوریتم Direct Preference Optimization (DPO) نسبت به انتخاب ضریب β مورد بررسی قرار گرفت و یک مکانیزم تنظیم تطبیقی برای این ضریب ارائه شد. نتایج تجربی فصل ۴ به‌طور روشن نشان دادند که انتخاب β در نسخه‌ی استاندارد DPO نقش تعیین‌کننده‌ای در پایداری آموزش دارد و افزایش آن لزوماً منجر به بهبود یادگیری ترجیحات انسانی نمی‌شود. آزمایش‌های انجام‌شده با سه مقدار ثابت $\beta = 0.05$، $\beta = 0.10$ و $\beta = 0.30$ نشان داد که با افزایش β، مقدار نهایی تابع زیان از حدود 0.74 به حدود 0.92 افزایش یافته و reward margin به مقادیر منفی‌تری میل کرده است. به‌طور هم‌زمان، نُرم گرادیان‌ها رشد قابل‌توجهی داشته و از مقادیری در حدود 30 در $\beta = 0.05$ به بیش از 200 در $\beta = 0.30$ رسیده است. این رفتار نشان‌دهنده‌ی اعمال فشار بیش‌ازحد ترجیحی و افزایش خطر ناپایداری عددی در فرآیند آموزش است، در حالی که دقت پاداش (reward accuracy) در تمامی مقادیر β تقریباً ثابت باقی مانده است. در مقابل، روش پیشنهادی Adaptive-β DPO با تنظیم پویا و کنترل‌شده‌ی ضریب β، امکان اعمال محافظه‌کارانه‌ی ترجیحات انسانی در مراحل ابتدایی آموزش و افزایش تدریجی شدت آن در مراحل پایدارتر را فراهم می‌کند. تحلیل لاگ‌های آموزشی نشان داد که در این روش، β از مقادیر پایین شروع شده و به‌صورت تدریجی به کران بالایی تعریف‌شده نزدیک می‌شود، بدون آنکه نوسانات شدید در تابع زیان یا نُرم گرادیان‌ها مشاهده شود. این نتایج نشان می‌دهند که تنظیم تطبیقی β می‌تواند پایداری آموزش را بهبود داده و وابستگی DPO به تنظیم دستی این پارامتر حساس را کاهش دهد.

5.2. دستاوردهای اصلی پژوهش

مهم‌ترین دستاورد این پژوهش، ارائه‌ی یک مکانیزم تنظیم تطبیقی برای ضریب β در چارچوب DPO است که بدون تغییر ساختار اصلی تابع زیان و بدون نیاز به مدل پاداش مجزا، پایداری فرآیند آموزش را بهبود می‌دهد. این روش با استفاده از سیگنال‌های درون‌مدلی نظیر اختلاف لگاریتم احتمال پاسخ ترجیحی و ناترجیحی، شدت اعمال ترجیحات انسانی را به‌صورت پویا تنظیم می‌کند. دستاورد دوم، ارائه‌ی یک تحلیل تجربی شفاف از حساسیت DPO نسبت به انتخاب β است. نتایج نشان می‌دهند که افزایش β می‌تواند منجر به افزایش شدید نُرم گرادیان‌ها و ناپایداری عددی شود، در حالی که بهبود معناداری در معیارهای مبتنی بر پاداش مشاهده نمی‌شود. این تحلیل، ضرورت استفاده از روش‌های تطبیقی به‌جای تنظیم دستی β را برجسته می‌کند. در نهایت، این پژوهش چارچوبی تجربی و قابل تفسیر برای تحلیل پایداری آموزش در الگوریتم‌های مبتنی بر ترجیح انسانی ارائه می‌دهد که می‌تواند در توسعه و ارزیابی روش‌های هم‌ترازی آینده مورد استفاده قرار گیرد.

5.3. محدودیت های پژوهش

نتایج ارائه‌شده در این پایان‌نامه در مقیاس محدود و با هدف اعتبارسنجی پیاده‌سازی و تحلیل رفتار دینامیکی آموزش به‌دست آمده‌اند. تمامی آزمایش‌ها با استفاده از مدل GPT-2 و بر روی زیرمجموعه‌ای کوچک از داده‌های Helpful–Harmless انجام شده‌اند و ارزیابی کیفی نهایی پاسخ‌ها (مانند win-rate انسانی یا خودکار) در این مرحله مدنظر نبوده است. همچنین، معیار واگرایی KL در این پژوهش به‌صورت یک شاخص جانشین ساده اندازه‌گیری شده و تحلیل دقیق انحراف مدل از سیاست مرجع می‌تواند در نسخه‌های تکمیلی با روش‌های دقیق‌تر انجام شود. از این رو، نتایج ارائه‌شده بیشتر بر پایداری آموزش و رفتار الگوریتم تمرکز دارند تا قضاوت نهایی درباره‌ی کیفیت خروجی مدل.

5.4. مسیر های آینده

به‌عنوان مسیرهای آینده، می‌توان روش پیشنهادی Adaptive-β DPO را در مقیاس بزرگ‌تر و بر روی مدل‌های زبانی پیشرفته‌تر، به‌ویژه مدل‌های تخصصی کدنویس، مورد ارزیابی قرار داد. استفاده از مجموعه‌داده‌های ترجیحی بزرگ‌تر و

متنوع‌تر می‌تواند امکان تحلیل دقیق‌تر رفتار این مکانیزم در سناریوهای واقعی‌تر را فراهم کند. علاوه بر این، به‌کارگیری معیارهای کیفی پیشرفته‌تر نظیر win-rate انسانی، reward margin تجمیعی و اندازه‌گیری دقیق واگرایی KL نسبت به سیاست مرجع می‌تواند ارزیابی جامع‌تری از اثرگذاری روش پیشنهادی ارائه دهد. در نهایت، تعمیم ایده‌ی تنظیم تطبیقی شدت ترجیح به سایر الگوریتم‌های هم‌ترازی مبتنی بر ترجیح می‌تواند مسیر پژوهشی ارزشمندی برای مطالعات آینده باشد.

[1, 5, 7-31]

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.