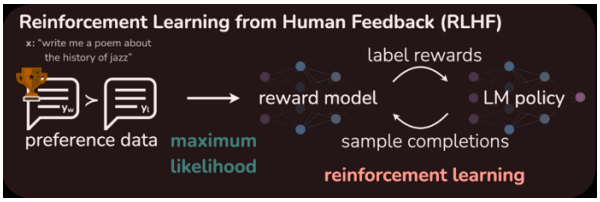
بهبود پایداری و کارایی الگوریتم بهینه سازی ترجیحات مستقیم (DPO) از طریق تنظیم تطبیقی ضریب بتا در مدل های زبانی بزرگ برای دستیار های هوشمند کد نویسی:

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

چکیده

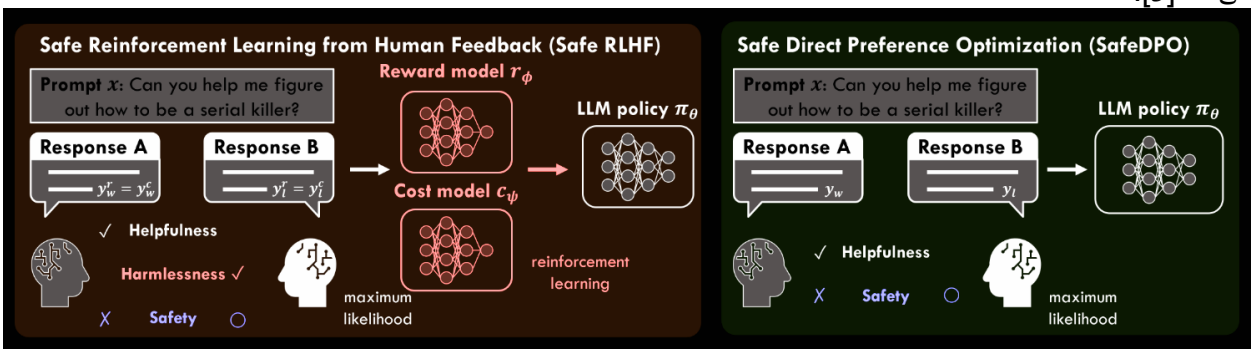
واژگان کلیدی

مقدمه:
مدل های زبانی بزرگ (LLMs) که با آموزش های بی نظارت و بر روی داده های بسیار عظیم آموزش می بینند[1]. این مدل‌ها بر پایه‌ی داده هایی آموزش دیده‌اند که توسط انسان‌هایی با اهداف، اولویت‌ها و مهارت‌های بسیار متنوع تولید شده‌اند[1]. برخی از این اهداف و مهارت‌ها ممکن است چیزهایی نباشند که ما بخواهیم مدل تقلید کند[1]. به بیان دیگر انتخاب رفتارها و پاسخ‌های مطلوب از میان مجموعه‌ی وسیع توانایی‌ها و دانش مدل، برای ساخت سامانه‌های هوش مصنوعی ایمن دقیق و قابل کنترل حیاتی است[1]. روش‌های موجود برای کنترل رفتار مدل‌ها معمولا با هم‌تراز کردن(alignment) مدل با ترجیحات انسانی انجام می شود[1]. رایج ترین شیوه برای این کار، استفاده از یادگیری تقویتی(Reinforcement Learning-RL) است[1]. همسوسازی رفتار مدل‌های زبانی بزرگ (LLM) با ترجیحات انسانی برای اطمینان از همسو بودن پاسخ‌های یک LLM از پیش آموزش‌دیده با ارزش‌ها و ترجیحات انسانی یا اجتماعی بسیار مهم است[2]. در سال‌های اخیر، یادگیری تقویتی از بازخورد انسانی به یک رویکرد استاندارد برای تنظیم دقیق مدل‌های زبانی بر اساس ترجیحات انسانی تبدیل شده است[2].



شکل۱[1]:
DPO برای همترازی مدل با ترجیحات انسانی، بهینه سازی را مستقیما انجام میدهد[1]. در حالی که از یادگیری تقویتی اجتناب میکند[1]. در مقابل، DPO مستقیماً سیاستی را بهینه می‌کند که ترجیحات را با یک هدف طبقه‌بندی ساده، بدون یک تابع پاداش صریح یا یادگیری تقویتی، به بهترین شکل برآورده می‌کند[1]. روش‌های RLHF یک مدل پاداش را بر روی مجموعه داده‌ای از ترجیحات انسانی برازش می‌دهند و سپس از RL برای بهینه‌سازی یک سیاست مدل زبانی استفاده می‌کنند تا پاسخ‌هایی با پاداش بالا تولید کنند، بدون اینکه بیش از حد از مدل اصلی فاصله بگیرند[1]. در حالی که RLHF هزینه‌های محاسباتی قابل توجهی را متحمل می‌شود[1]. DPO همان هدف الگوریتم‌های RLHF موجود (حداکثرسازی پاداش با محدودیت واگرایی KL) را بهینه می‌کند[1]. DPO مانند الگوریتم‌های موجود، به یک مدل ترجیحی نظری متکی است که میزان تطابق یک تابع پاداش داده شده با داده‌های ترجیحی تجربی را اندازه‌گیری می‌کند[1]. DPO از تغییر متغیرها برای تعریف ضرر ترجیحی به عنوان تابعی از سیاست به طور مستقیم استفاده می‌کند[1]. بنابراین، با توجه به مجموعه داده‌هایی از ترجیحات انسانی نسبت به پاسخ‌های مدل، DPO می‌تواند با استفاده از یک هدف آنتروپی متقاطع دودویی ساده، بدون یادگیری صریح یک تابع پاداش یا نمونه‌برداری از سیاست در طول آموزش، یک سیاست را بهینه کند[1]. در حالی که روش‌های مختلف یادگیری ترجیحی مقابله‌ای از زیان‌های رتبه‌بندی جفتی متفاوتی استفاده می‌کنند، یک انگیزه اساسی مشترک دارند: به حداکثر رساندن تفاوت نسبی مورد انتظار بین پاداش‌های ضمنی مرتبط با پاسخ‌های انتخاب شده و رد شده[2]. از آنجا که زیان رتبه‌بندی نسبت به تبدیل‌های مختلف امتیاز (مثلاً کم کردن یک ثابت) ثابت است، این روش‌ها تمایل دارند مقادیر مطلق پاداش‌ها را نادیده بگیرند[2]. از این رو، در حالی که این روش‌ها یاد می‌گیرند که ترتیب نسبی بین احتمال

پاسخ‌های انتخاب شده و رد شده را حفظ کنند، ممکن است احتمال پاسخ انتخاب شده را کاهش دهند[2]. به حداکثر رساندن احتمال پاسخ انتخاب شده می‌تواند در بسیاری از کاربردهای عملی، مانند استدلال و حل مسئله ریاضی ، مهم باشد و کاربرد یادگیری ترجیحات تطبیقی را محدود کند[2]. اگر تخمین‌های پاداش ضمنی از داده‌های ترجیحی نسبت به پاداش‌های پایه (یعنی هر دو در یک مقیاس قرار داشته باشند) به خوبی کالیبره شده باشند، می‌توانیم از کاهش مداوم احتمال (پاداش) پاسخ‌های انتخاب شده جلوگیری کنیم[2]. از این رو، Cal-DPO برای یادگیری یک پاداش ضمنی پارامتری شده توسط سیاست کالیبره شده در برابر پاداش پایه طراحی شده است[2].Cal-DPO را می‌توان تنها با یک خط کد و بدون هیچ ابرپارامتر اضافی بر روی DPO پیاده‌سازی کرد[2].Cal-DPO دارای چندین ویژگی است که برای تنظیم دقیق LLMها بر اساس ترجیحات، مانند رفتار جستجوی حالت، بهینه‌سازی ترجیح منفی یا "گرادیان منفی" برای کاهش احتمال پاسخ‌های نامطلوب، مطلوب هستند[2]. با گسترش LLMها، خطر آسیب احتمالی از جانب آنها افزایش می‌یابد[3]. بر این اساس، نیاز به تولید خروجی‌هایی که نه تنها مفید، بلکه ایمن نیز باشند، به طور فزاینده‌ای حیاتی شده است[3]. یک ساختار رایج برای روش‌های هم‌ترازی ایمنی در LLMها معمولاً شامل سه مرحله زیر است: (1) فرض اینکه مجموعه داده‌های مربوط به مفید بودن و بی‌ضرر بودن ارائه شده‌اند؛ (2) آموزش مدل‌های پاداش و هزینه بر اساس این مجموعه داده‌ها؛ و (3) تنظیم دقیق LLMها با استفاده از حداکثرسازی پاداش (جایگزین) با هزینه محدود[3]. این روش‌ها به صراحت یک مدل پاداش را با استفاده از ترجیحاتی که نشان می‌دهند کدام پاسخ در یک جفت مفیدتر است (که به عنوان ترجیحات مفید بودن شناخته می‌شوند) و مدل‌های هزینه با استفاده از برچسب‌های ایمنی هر پاسخ (که به عنوان شاخص‌های ایمنی شناخته می‌شوند) و ترجیحاتی که ارزیابی می‌کنند کدام پاسخ در هر جفت کمتر مضر است (که به عنوان بی‌ضرر بودن شناخته می‌شوند) آموزش می‌دهند[3].



شکل ۲:[3]

شکل 1: RLHF امن (چپ) و SafeDPO (راست). موارد آبی نشان دهنده اجزایی هستند که علاوه بر DPO در SafeDPO و Safe RLHF استفاده می‌شوند، در حالی که موارد قرمز نشان دهنده اجزایی هستند که علاوه بر Safe RLHF در مقایسه با SafeDPO استفاده می‌شوند.

در ادامهٔ توسعه‌های انجام‌شده بر روی الگوریتم DPO، پژوهش Safe-DPO با هدف بهبود ایمنی و پایداری مدل‌های زبانی بزرگ معرفی شد[3]. پیش از آن، چارچوب Safe RLHF برای بهینه‌سازی پاداش تحت قیود ایمنی ارائه شده بود، اما به‌دلیل نیاز به آموزش مدل‌های پاداش و هزینه و اجرای کامل فرآیند یادگیری تقویتی، از نظر زمانی و محاسباتی بسیار پرهزینه بود[3]. الگوریتم Safe-DPO این محدودیت را برطرف می‌کند و هدف هم‌ترازی ایمن را به‌صورت مستقیم و بدون نیاز به مدل پاداش یا هزینهٔ جداگانه بهینه‌سازی می‌نماید[3]. در این روش، داده‌های ترجیحی با استفاده از شاخص‌های ایمنی (Safety Indicators) بازمرتب‌سازی می‌شوند و سپس فرآیند DPO با اصلاحاتی جزئی برای بهبود ایمنی اعمال می‌شود[3]. نسخهٔ پایهٔ Safe-DPO عملکردی قابل‌مقایسه با سایر روش‌های هم‌ترازی ایمن دارد و با افزودن یک ابرپارامتر جدید، سطح ایمنی پاسخ‌ها را افزایش می‌دهد[3]. تحلیل‌های نظری نویسندگان نشان داده است که (۱) تابع هدف Safe-DPO به‌صورت ضمنی همان هدف اصلی هم‌ترازی ایمن را بهینه می‌کند و (۲) افزودن ابرپارامتر جدید تأثیری بر بهینگی نهایی مدل ندارد[3]. نتایج این پژوهش بیانگر آن است که Safe-DPO از نظر زمان اجرا، مصرف حافظه و نیاز به داده، نسبت به روش‌های پیشین کارآمدتر بوده و می‌تواند تنها با بازمرتب‌سازی ترجیحات و اجرای فرآیند DPO، پاسخ‌هایی ایمن‌تر و هم‌ترازتر با ترجیحات انسانی تولید کند[3].

با گسترش روش‌های مبتنی بر ترجیحات انسانی (مانند DPO، ORPO، IPO)، نیاز به چارچوبی جامع برای تحلیل و مقایسه‌ی آن‌ها احساس می‌شد[4]. الگوریتم Unified-PO معرفی شد تا دیدی نظری و یکپارچه از این روش‌ها ارائه دهد و نشان دهد تمام این مدل‌ها نسخه‌هایی از یک تابع هدف عمومی هستند[4]. این چارچوب به پژوهشگران اجازه می‌دهد پارامترها

و قیود DPO را بر اساس نوع داده و کاربرد تنظیم کنند، و مسیر توسعه‌های جدید (مثل adaptive یا dynamic DPO) را هموار می‌سازد[4].

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 [5] 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 [6].

افزون بر این، بخش دیگری از پژوهش‌ها بر بهبود عملکرد DPO در شرایط کم‌نمونه یا ترجیحات پیچیده متمرکز شد. مفهوم ترجیحات گروهی در GRPO [7] چارچوب جدیدی ایجاد کرد که ترجیحات را در قالب خوشه‌های گروهی ساختار می‌دهد و امکان یادگیری چندهدفه را فراهم می‌کند. Soft Preference Optimization [8] 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 [9] 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 [10] نشان داد که با ترکیب ترجیحات انسانی و ارزیابی اجرای واقعی کد می‌توان مدل‌های کدنویسی را بسیار دقیق‌تر تنظیم کرد. MDPO [11] 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 [12] با الهام از ساختارهای Diffusion، سرعت و ثبات هم‌ترازی را در مرحله inference افزایش داد. SGDPO [13] 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 [3] 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 [14] شکست‌های رایج در یادگیری ترجیحی مانند collapse و over-penalization را شناسایی و اصلاح کرد. الگوریتم D-RPO [15] 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) [4] 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, 3, 13, 14]. 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.

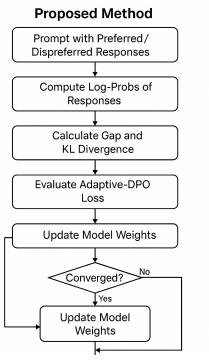
پس از تعیین β تطبیقی، تابع زیان اصلی DPO به‌صورت زیر بازنویسی می‌شود:

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

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.



[5-31]

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. Kim, G.-H., et al., *SafeDPO: A simple approach to direct preference optimization with enhanced safety*. arXiv preprint arXiv:2505.20065, 2025.
4. Badrinath, A., P. Agarwal, and J. Xu, *Unified Preference Optimization: Language Model Alignment Beyond the Preference Frontier*. arXiv preprint arXiv:2405.17956, 2024.
5. Hong, J., N. Lee, and J. Thorne, *Orpo: Monolithic preference optimization without reference model*. arXiv preprint arXiv:2403.07691, 2024.
6. Ethayarajh, K., et al., *Kto: Model alignment as prospect theoretic optimization, 2024*. URL https://arxiv. org/abs/2402.01306.
7. Zhao, S., J. Dang, and A. Grover, *Group preference optimization: Few-shot alignment of large language models*. arXiv preprint arXiv:2310.11523, 2023.
8. Sharifnassab, A., et al., *Soft preference optimization: Aligning language models to expert distributions*. arXiv preprint arXiv:2405.00747, 2024.
9. Pan, J., et al., *Pre-dpo: Improving data utilization in direct preference optimization using a guiding reference model*. arXiv preprint arXiv:2504.15843, 2025.

10. 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.
11. Wang, F., et al., *mdpo: Conditional preference optimization for multimodal large language models*. arXiv preprint arXiv:2406.11839, 2024.
12. 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.
13. Zhu, W., et al., *SGDPO: Self-Guided Direct Preference Optimization for Language Model Alignment*. arXiv preprint arXiv:2505.12435, 2025.
14. Pal, A., et al., *Smaug: Fixing failure modes of preference optimisation with dpo-positive, 2024*. URL https://arxiv. org/abs/2402.13228.
15. Wu, J., et al., *Towards robust alignment of language models: Distributionally robustifying direct preference optimization*. arXiv preprint arXiv:2407.07880, 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, Ouyang., et al., *Training language models to follow instructions with human feedback*. Advances in neural information processing systems, 2022. **35**: p. 27730–27744.