ارائه‌ی یک روش برای بهینه سازی ترجیحات انسانی مستقیم در توسعه سامانه های هوش مصنوعی قابل اعتماد

عارف گنجائی ساری
گروه مهندسی کامپیوتر، دانشکده فنی و مهندسی، دانشگاه آزاد اسلامی واحد غرب، تهران، 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 (Large Language Models) در سال‌های اخیر به هسته‌ی اصلی بسیاری از سامانه‌های هوش مصنوعی پیشرفته تبدیل شده‌اند و توانسته‌اند در وظایفی نظیر مکالمه، استدلال، تولید کد و تحلیل متون عملکرد چشمگیری از خود نشان دهند. با این حال، آموزش این مدل‌ها عمدتاً بر پایه‌ی داده‌های عظیم و ناهمگون انسانی انجام می‌شود؛ داده‌هایی که ترکیبی از رفتارهای مطلوب، خطاهای انسانی، سوگیری‌ها و پاسخ‌های نامطلوب را در خود دارند. از این رو، همترازسازی رفتار مدل‌های زبانی با ترجیحات، ارزش‌ها و انتظارات انسانی به یکی از چالش‌های بنیادین در توسعه‌ی LLMها تبدیل شده است. نخستین چارچوب موفق در این مسیر، یادگیری تقویتی از بازخورد انسانی (RLHF) بود که نشان داد می‌توان با استفاده از ترجیحات انسانی، رفتار مدل را به‌صورت مؤثر هدایت کرد [1]. با وجود موفقیت RLHF، پیچیدگی محاسباتی بالا، نیاز به آموزش مدل پاداش مجزا و ناپایداری ذاتی الگوریتم‌های یادگیری تقویتی، محدودیت‌های جدی این رویکرد را آشکار ساخت. در پاسخ به این چالش‌ها، الگوریتم Direct Preference Optimization (DPO) معرفی شد که با یک بازنویسی هوشمندانه از هدف RLHF، نیاز به مدل پاداش و حلقه‌ی یادگیری تقویتی را به‌طور کامل حذف می‌کند و همترازی را مستقیماً از طریق یک تابع زیان مبتنی بر ترجیحات انسانی انجام می‌دهد [2] .پس از DPO، پژوهش‌ها نشان دادند که اگرچه این الگوریتم ساده و کارآمد است، اما پاداش‌های ضمنی استخراج‌شده از آن لزوماً با مقیاس پاداش پایه همخوانی ندارند. Cal-DPO با هدف کالیبره‌کردن این پاداش‌های ضمنی معرفی شد و نشان داد که می‌توان بدون افزودن پیچیدگی قابل‌توجه، پایداری یادگیری و احتمال پاسخ‌های منتخب را بهبود بخشید [3]. همزمان، مسئله‌ی ایمنی نیز به‌عنوان یکی از دغدغه‌های اصلی در همترازی LLMها مطرح شد. Safe-DPO نشان داد که می‌توان با بازمرتب‌سازی ترجیحات بر اساس شاخص‌های ایمنی، کنترل رفتارهای مضر را بدون استفاده از یادگیری تقویتی یا مدل‌های هزینه‌ی جداگانه محقق کرد [4]. در ادامه‌ی این مسیر، تلاش‌هایی برای ساده‌سازی بیشتر چارچوب DPO انجام شد. ORPO با حذف سیاست مرجع، فرآیند همترازی را سبکتر کرد [5] و Unified Preference Optimization نشان داد که بسیاری از روش‌های مبتنی بر ترجیح را می‌توان به‌عنوان حالت‌های خاصی از یک چارچوب نظری یکپارچه تحلیل کرد [6]. از سوی دیگر، مسئله‌ی کارایی داده نیز مورد توجه قرار گرفت؛ Pre-DPO نشان داد که استفاده‌ی هوشمندانه از مدل مرجع راهنما می‌تواند بهره‌وری داده‌های ترجیحی را افزایش دهد [7] و SGDPO مفهوم یادگیری خودراهبری را برای کاهش وابستگی به بازخورد انسانی مطرح کرد [8]. برخی پژوهش‌ها نیز بر بهبود پایداری زمانی و کارایی استنتاج تمرکز داشتند. DiffPO با الهام از مدل‌های Diffusion، همترازی در زمان استنتاج را سریع‌تر و پایدارتر کرد [9]. در بعد نظری، KTO با بهره‌گیری از نظریه‌ی چشم‌انداز، نشان داد که می‌توان حساسیت به ریسک و مطلوبیت واقعی انسان را مستقیماً در فرآیند همترازی وارد کرد [10]. همچنین D-RPO با نگاه توزیعی، پایداری DPO را در داده‌های نویزدار افزایش داد [11] و MoE-DPO با استفاده از معماری Mixture-of-Experts توانست ترجیحات پیچیده را بهتر مدل‌سازی کند [12]. همزمان، رویکردهایی مانند Group Preference Optimization [13] و Soft Preference Optimization [14]] تلاش کردند همترازی را در سناریوهای چندهدفه و ترجیحات نرم بهبود دهند. در حوزه‌ی کاربردهای تخصصی، PLUM نشان داد که ترکیب ترجیحات انسانی با اجرای واقعی کد می‌تواند همترازی مدل‌های کدنویس را به‌طور معناداری ارتقا دهد [15] و MDPO چارچوب DPO را به مدل‌های چندوجهی گسترش داد [16]. روش‌هایی مانند Smaug [17] و پژوهش‌های مرتبط با بهینه‌سازی کارآمد همترازی [18] نیز به بررسی و رفع حالت‌های شکست در روش‌های ترجیح‌محور پرداختند. در سال‌های اخیر، توجه به ترجیحات متنوع، چندهدفه و تطبیقی نیز افزایش یافته است. Auto-Weighted Group RPO [19]، Contrastive Preference Optimization [20] و Self-Rewarding Language Models [21] هر یک از زاویه‌ای متفاوت تلاش کرده‌اند محدودیت‌های همترازی سنتی را کاهش دهند. همچنین بررسی نقش سیاست مرجع در DPO [22]، مطالعات مروری جامع [23–26]، روش‌های شتاب‌داده‌شده [27]، ترجیحات متنوع [28] و چارچوب‌های آگاه از پاداش [29] نشان می‌دهند که همترازی ترجیح‌محور همچنان حوزه‌ای فعال و در حال تکامل است. در نهایت، روش‌های تطبیقی در حوزه‌های چندوجهی [30] و مهندسی نمایش [31] نیز اهمیت تنظیم پویا رفتار مدل را بیش از پیش برجسته کرده‌اند. با وجود این پیشرفت‌های گسترده، یک محدودیت بنیادین در تمامی این روش‌ها مشترک است: کنترل شدت اعمال ترجیحات انسانی همواره بر پایه‌ی یک ضریب ثابت β انجام می‌شود. این در حالی است که β نقشی تعیین‌کننده در مقیاس گرادیان‌ها، پایداری همگرایی و میزان واگرایی KL نسبت به سیاست مرجع دارد. در همین راستا، این پژوهش با معرفی الگوریتم Adaptive-β DPO تلاش می‌کند تا با تنظیم تطبیقی β بر اساس سیگنال‌های درون‌مدلی و شرایط آموزشی، پایداری آموزش و کارایی همترازی را به‌طور همزمان بهبود دهد.

پیشینه تحقیق:

با گسترش مدل‌های زبانی در سال‌های اخیر، روش‌های متعددی برای هم‌ترازی مدل‌های زبانی با ترجیحات انسانی پیشنهاد شده‌اند. جدول (۱) مروری ساختاریافته بر مهم‌ترین این روش‌ها ارائه می‌دهد. پیش از ارائه جدول، در ادامه روند تکامل این روش‌ها به‌صورت 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 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]

جدول ۱ ـ مروری مقایسه‌ای بر روش‌های مبتنی بر ترجیح انسانی و چالش‌های باقی‌مانده

ردیف	منبع	سال	عنوان مقاله	منبع انتشار	چکیده کوتاه	نتایج عددی	الگوریتم	مدل/چارچوب	چالش‌ها
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 پاسخ‌های منتخب	KL ۲۵٪↓، ۸٪↑ پایداری	Calibrated DPO	GPT-J, HH	β همچنان ثابت؛ پایداری محدود در داده نویزدار
3	[4]	2025	Safe-DPO	arXiv 2025	بازمرتب‌سازی ترجیحات با شاخص‌های ایمنی بدون نیاز به مدل پاداش	۲۳٪↓ خطای پاداش در داده‌های نویزدار	Safe-DPO	Reddit + HH	ایمنی↑ ولی β تطبیقی ندارد

4	[7]	2025	Pre-DPO	arXiv 2025	استفاده از مدل مرجع راهنما برای بهبود کارایی داده	Data ۱۵٪↑ ،Efficiency loss۵٪↓	Pre-DPO	Anthropic-HH	βثابت؛ حساسیت به توزیع داده
5	[8]	2025	SGDPO	ACL 2025	یادگیری خودراهبری برای کاهش نیاز به داده انسانی	win-۹٪↑ rate، ۱۲٪↑ stability	Self-Guided DPO	TL;DR, HH	βثابت؛ 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 بالا 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

بررسی تطبیقی روش‌های ارائه‌شده در جدول(۱) نشان می‌دهد که اگرچه الگوریتم DPO نقطه‌ی عطفی در ساده‌سازی فرآیند هم‌ترازی مدل‌های زبانی بوده است، اما مسئله‌ی پایداری آموزش همچنان به‌عنوان یک چالش بنیادین در اغلب نسخه‌ها و توسعه‌های بعدی آن باقی مانده است. تقریباً تمامی روش‌های مبتنی بر ترجیحات انسانی—از جمله DPO، Cal-DPO، Safe-DPO، ORPO، KTO، D-RPO و نسخه‌های پیشرفته‌تر نظیر MoE-DPO—برای کنترل شدت اعمال ترجیحات انسانی و میزان انحراف سیاست مدل از سیاست مرجع، به استفاده از یک ضریب ثابت β متکی هستند. این در حالی است که نتایج عددی گزارش‌شده در این مطالعات نشان می‌دهد انتخاب مقدار β تأثیر مستقیمی بر نُرم گرادیان‌ها، رفتار همگرایی، میزان افزایش یا کاهش واگرایی KL و در نهایت پایداری آموزش دارد. اگرچه برخی پژوهش‌ها تلاش کرده‌اند با اصلاح تابع هدف (مانند KTO)، بازمرتب‌سازی داده‌ها (مانند Safe-DPO) یا بهبود کارایی داده (مانند Pre-DPO) بخشی از مشکلات عملی DPO را کاهش دهند، اما هیچیک به‌طور مستقیم به مسئله‌ی تنظیم پویا و مرحله‌به‌مرحله‌ی β نپرداخته‌اند. در اغلب این روش‌ها، مقدار β پیش از آموزش انتخاب شده و در طول کل فرآیند یادگیری بدون تغییر باقی می‌ماند. این رویکرد ثابت، با ماهیت پویای فرآیند یادگیری ترجیحی—که شامل نمونه‌های ساده و دشوار، مراحل ناپایدار اولیه و مراحل پایدار پایانی است—سازگاری کامل ندارد. نتایج مطالعات موجود به‌طور ضمنی نشان می‌دهند که β بزرگ می‌تواند منجر به فشار بیش‌ازحد ترجیحی، رشد شدید نُرم گرادیان‌ها و افزایش ناخواسته‌ی KL شود، در حالی که β کوچک ممکن است یادگیری ترجیحات انسانی را تضعیف کرده و به همگرایی کند یا ناکافی منجر شود. با این حال، در ادبیات موجود مکانیزمی داده‌محور و خودتطبیقی برای تنظیم β ارائه نشده است که بتواند به‌صورت همزمان عدم‌قطعیت مدل نسبت به ترجیحات انسانی و میزان فاصله‌ی آن از سیاست مرجع را در نظر بگیرد. این خلأ پژوهشی نشان می‌دهد که بهبود پایداری آموزش و کنترل مؤثر واگرایی KL در الگوریتم‌های مبتنی بر ترجیح انسانی، نیازمند رویکردی است که پارامتر β را نه به‌عنوان یک ابرپارامتر ثابت، بلکه به‌عنوان یک متغیر پویا و وابسته به وضعیت آموزش در نظر بگیرد. بر همین اساس، در این پژوهش روش Adaptive-β DPO پیشنهاد می‌شود که در آن مقدار β به‌صورت تطبیقی و بر اساس سیگنال‌های درون‌مدلی تنظیم می‌گردد تا مدل بتواند در نمونه‌های دشوار یادگیری قوی‌تری داشته باشد و در شرایط ناپایدار، رفتار محافظه‌کارانه‌تری برای حفظ پایداری اتخاذ کند.

روش پیشنهادی Adaptive-β (DPO):

الگوریتم Direct Preference Optimization (DPO) به عنوان 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. در این چارچوب، شدت اعمال ترجیحات انسانی توسط پارامتر β کنترل می‌شود. در نسخه‌ی استاندارد DPO، 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. در روش پیشنهادی Adaptive-β DPO، این محدودیت با معرفی یک مکانیزم تنظیم تطبیقی برای ضریب β برطرف می‌شود. ایده‌ی اصلی روش بر این مبنا استوار است که شدت اعمال ترجیحات انسانی باید با وضعیت لحظه‌ای مدل در طول آموزش سازگار باشد، نه آنکه به‌صورت یکنواخت و ثابت اعمال شود. فرض می‌شود مجموعه‌داده‌ی آموزشی شامل نمونه‌های ترجیحی به‌صورت سه‌تایی (x, y_w, y_l) باشد که در آن x پرامپت ورودی، y_l پاسخ ترجیحی و y_w پاسخ ناترجیحی است. سیاست فعلی مدل با π_θ و سیاست مرجع با π_{ref} نمایش داده می‌شود. برای هر نمونه، اختلاف لگاریتمی احتمال پاسخ‌ها تحت سیاست فعلی به‌صورت زیر تعریف می‌شود:

$$\Delta_\theta(x, y_w, y_l) = \log \pi_\theta(y_w|x) - \log \pi_\theta(y_l|x) \qquad \text{(رابطه ۱)}$$

این کمیت میزان تمایز مدل میان پاسخ ترجیحی و ناترجیحی را نشان می‌دهد. مقادیر بزرگ و مثبت بیانگر اطمینان بالای مدل نسبت به ترجیح انسانی هستند، در حالی که مقادیر کوچک یا منفی نشان‌دهنده‌ی عدم‌قطعیت مدل و دشواری نمونه می‌باشند. ازاین‌رو، π_θ به‌عنوان یک سیگنال درون‌مدلی برای سنجش میزان اطمینان مدل نسبت به ترجیحات انسانی مورد استفاده قرار می‌گیرد. تابع زیان استاندارد DPO به‌صورت زیر تعریف می‌شود:

$$\mathcal{L}_{DPO}(\theta) = -\mathbb{E}[\log \sigma(\beta(\Delta_\theta - \Delta_{ref}))] \qquad \text{(رابطه ۲)}$$

که در آن $\sigma(\cdot)$ تابع سیگموید و β ضریب شدت اعمال ترجیحات انسانی است. این ضریب مستقیماً مقیاس گرادیان‌ها را کنترل می‌کند و نقش تعیین‌کننده‌ای در پایداری همگرایی دارد. علاوه بر سیگنال عدم‌قطعیت ترجیحی، برای کنترل انحراف مدل از رفتار اولیه، یک معیار جانشین برای واگرایی کولبک–لایبلر میان سیاست فعلی و سیاست مرجع در نظر گرفته می‌شود:

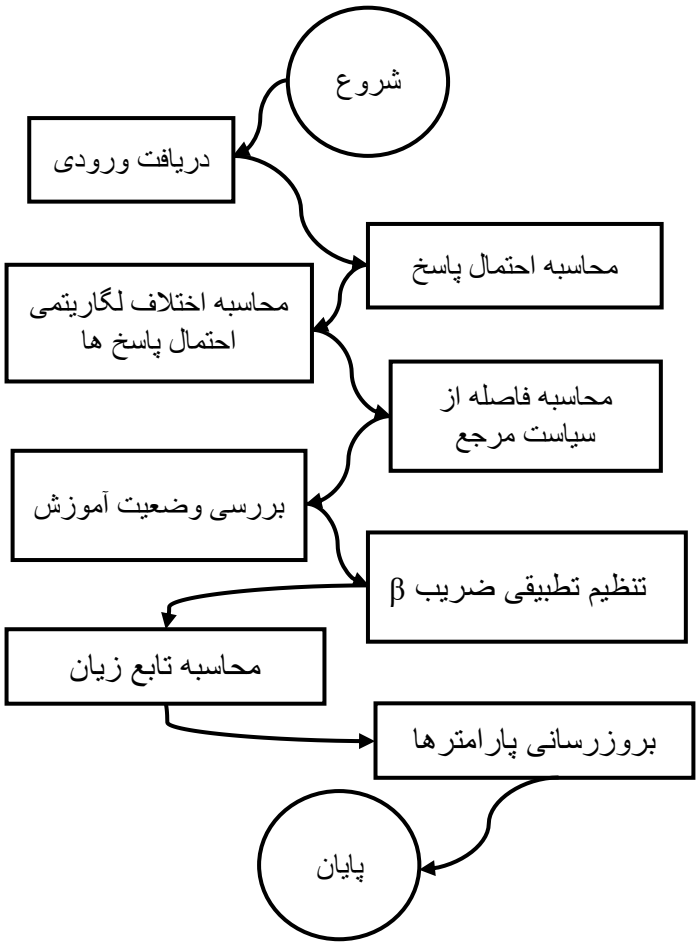
$$\widehat{D}_{KL} = D_{KL}(\pi_\theta - \pi_{ref}) \qquad \text{(رابطه ۳)}$$

این کمیت میزان فاصله‌ی مدل از سیاست مرجع را در طول فرآیند آموزش بازتاب می‌دهد و نقش یک سیگنال کنترلی برای جلوگیری از drift و ناپایداری عددی ایفا می‌کند. در Adaptive-β DPO، ضریب β به‌صورت پویا و بر اساس ترکیبی از این دو سیگنال به‌روزرسانی می‌شود:

$$\beta_{t+1} = clip(\beta_t + \eta_\beta g(\Delta_\theta) - \gamma . h(\widehat{D}_{KL}), \beta_{min}, \beta_{max} \qquad \text{(رابطه ۴)}$$

که در آن $g(\cdot)$ تابعی افزایشی از عدم‌قطعیت ترجیحی، $h(\cdot)$ تابعی افزایشی از فاصله از سیاست مرجع، و $[\beta_{min}, \beta_{max}]$ بازه‌ای ایمن برای جلوگیری از رفتارهای افراطی هستند. شهود این قاعده به این صورت است که در نمونه‌های دشوار (تمایز کم میان پاسخ‌ها)، مقدار β افزایش یافته و فشار یادگیری تقویت می‌شود، در حالی که در شرایط افزایش واگرایی نسبت به سیاست مرجع، مقدار β کاهش می‌یابد تا پایداری آموزش حفظ شود. مقدار به‌روزشده‌ی β مستقیماً در تابع زیان DPO مطابق رابطه (2) استفاده می‌شود و بدون تغییر در ساختار اصلی تابع هدف، شدت به‌روزرسانی گرادیان‌ها را کنترل می‌کند. در نتیجه، نمونه‌های دشوار با به‌روزرسانی‌های محافظه‌کارانه‌تر و نمونه‌های ساده‌تر با شدت یادگیری بالاتر پردازش می‌شوند. 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 فلوچارت ارائه‌شده نمایش داده شده است. به این ترتیب، Adaptive-β DPO را می‌توان به‌عنوان یک تعمیم پویا از DPO استاندارد در نظر گرفت که به همگرایی پایدارتر و کنترل بهتر واگرایی نسبت به سیاست مرجع منجر می‌شود.

شکل ۱ - فلوچارت گام های الگوریتم Adaptive DPO



در شکل ۱- مراحل کلی الگوریتم پیشنهادی Adaptive-β DPO نمایش داده شده است. فرآیند با دریافت داده‌های ترجیحی و محاسبه‌ی اختلاف لگاریتمی احتمال پاسخ‌های ترجیحی و ناترجیحی آغاز می‌شود. سپس میزان فاصله‌ی سیاست فعلی مدل از سیاست مرجع به‌عنوان یک سیگنال کنترلی محاسبه می‌گردد. بر اساس این دو سیگنال، مقدار ضریب β به‌صورت تطبیقی تنظیم شده و در تابع زیان DPO مورد استفاده قرار می‌گیرد. در نهایت، پارامترهای مدل به‌روزرسانی شده و این روند تا رسیدن به همگرایی ادامه می‌یابد. این سازوکار موجب پایداری بیشتر آموزش و کنترل بهتر واگرایی نسبت به سیاست مرجع می‌شود.

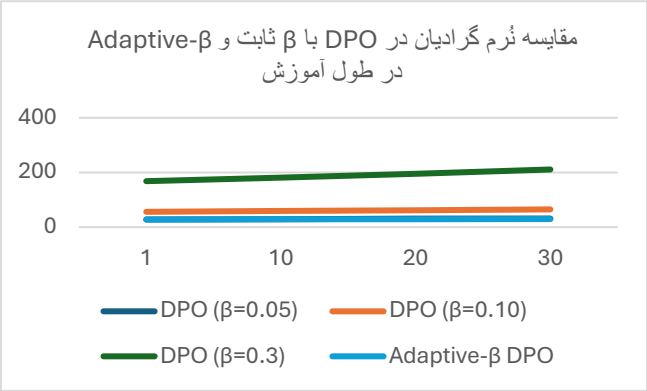
ارزیابی و نتایج:

در این بخش، عملکرد روش پیشنهادی Adaptive-β DPO در مقایسه با نسخه‌ی استاندارد Direct Preference Optimization با ضریب β ثابت مورد ارزیابی قرار می‌گیرد. هدف اصلی این ارزیابی، بررسی پایداری فرآیند آموزش، رفتار گرادیان‌ها و میزان کنترل انحراف مدل از سیاست مرجع در شرایط کنترل‌شده است، نه مقایسه‌ی نهایی کیفیت پاسخ‌ها. به همین منظور، از مجموعه‌داده‌ی ترجیحی استاندارد Helpful–Harmless(HH) استفاده شده است که شامل نمونه‌های سه‌تایی متشکل از پرامپت، پاسخ ترجیحی و پاسخ ناترجیحی می‌باشد و به‌طور گسترده در پژوهش‌های مبتنی بر DPO به‌کار رفته است. در تمامی آزمایش‌ها، ساختار داده‌ها، رویه‌ی پیش‌پردازش، معماری مدل پایه (GPT-2) و تنظیمات آموزشی در تمامی روش‌ها یکسان در نظر گرفته شده است تا مقایسه‌ای منصفانه و قابل تکرار حاصل شود. برای ارزیابی عملکرد، مجموعه‌ای از معیارهای مکمل مورد استفاده قرار گرفته‌اند که تمرکز اصلی آن‌ها بر پایداری عددی و رفتار یادگیری ترجیحات است. نخست، مقدار تابع زیان آموزشی و روند تغییرات آن در طول گام‌های یادگیری تحلیل شده است. 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. علاوه بر این، نُرم گرادیان‌ها به‌عنوان یکی از شاخص‌های کلیدی پایداری عددی آموزش مورد توجه قرار گرفته است، زیرا افزایش شدید این کمیت می‌تواند نشانه‌ای از فشار بیش‌ازحد ترجیحی و خطر ناپایداری در فرآیند بهینه‌سازی باشد. در روش پیشنهادی Adaptive-β DPO، رفتار دینامیکی ضریب β و تغییرات آن در طول آموزش نیز ثبت و تحلیل شده است تا ارتباط میان عدم‌قطعیت مدل و شدت اعمال ترجیحات انسانی بررسی شود. به‌منظور تحلیل حساسیت DPO نسبت به انتخاب مقدار β، نسخه‌ی استاندارد این الگوریتم با سه مقدار ثابت β = 0.05، β = 0.10 و β = 0.30 اجرا شده است. جدول (۲) خلاصه‌ای از نتایج حاصل از این آزمایش‌ها را ارائه می‌دهد که شامل مقدار نهایی تابع زیان، reward margin و نُرم گرادیان‌ها در پایان آموزش است. نتایج نشان می‌دهند که با افزایش مقدار β، مقدار نهایی تابع زیان افزایش یافته و reward margin به مقادیر منفی‌تری میل می‌کند. همچنین، نُرم گرادیان‌ها با افزایش β رشد چشمگیری داشته و در مقدار β = 0.30 به مقادیر بسیار بزرگی می‌رسد. این رفتار بیانگر آن است که افزایش β لزوماً منجر به بهبود یادگیری ترجیحات انسانی نمی‌شود و در عوض می‌تواند باعث اعمال فشار بیش‌ازحد ترجیحی و ناپایداری عددی آموزش گردد. در مقابل، دقت پاداش در تمامی مقادیر β تقریباً ثابت باقی مانده است که نشان می‌دهد تنظیم دستی β یک پارامتر حساس و بالقوه ناپایدار در DPO محسوب می‌شود.

جدول ۲. مقایسه‌ی DPO با مقادیر مختلف β ثابت

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

نمودار ۱. مقایسه‌ی DPO با مقادیر مختلف β ثابت و غیر ثابت



جدول ۳. مقایسه‌ی پژوهش Adaptive-β DPO با نتایج مقالات دیگر β ثابت

تمرکز اصلی	نیاز به مدل پاداش	پایداری گرادیان	کنترل KL	تنظیم β	روش
سادگی جایگزین RLHF	خیر	ناپایدار در β بزرگ	غیرمستقیم	ثابت	DPO [2]
کالیبراسیون پاداش	خیر	متوسط	جزئی	ثابت	Cal-DPO [3]
ایمنی خروجی	خیر	متوسط	جزئی	ثابت	Safe-DPO [4]
مقاومت به نویز	خیر	بهتر	جزئی	ثابت	D-RPO [11]
پایداری + کنترل KL	خیر	پایدار	فعال	تطبیقی	Adaptive-β DPO

همان‌طور که در این جدول مشاهده می‌شود، اغلب روش‌های پیشین از ضریب β ثابت استفاده می‌کنند و تمرکز آن‌ها بر جنبه‌هایی مانند کالیبراسیون پاداش، ایمنی یا مقاومت در برابر نویز بوده است. در مقابل، Adaptive-β DPO نخستین چارچوبی است که تنظیم ضریب β را به‌صورت پویا و داده‌محور انجام می‌دهد و به‌طور همزمان پایداری گرادیان‌ها و کنترل واگرایی KL نسبت به سیاست مرجع را هدف قرار می‌دهد، بدون آنکه نیاز به مدل پاداش مجزا یا استفاده از یادگیری تقویتی داشته باشد. 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 نشان‌دهنده‌ی اعمال محافظه‌کارانه‌ی ترجیحات انسانی در مراحل اولیه و افزایش تدریجی فشار ترجیحی در مراحل پایدارتر آموزش است. بررسی مقادیر اختلاف لگاریتمی احتمال پاسخ‌ها نشان می‌دهد که این کمیت در طول آموزش دارای نوسانات قابل‌توجه و حتی مقادیر منفی در برخی گام‌هاست که بیانگر وجود نمونه‌هایی با عدم‌قطعیت بالا نسبت به ترجیح انسانی است. مکانیزم Adaptive-β در چنین شرایطی، 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 نشان‌دهنده‌ی کنترل انحراف مدل از سیاست مرجع در طول آموزش می‌باشد. در مجموع، نتایج تجربی نشان می‌دهند که Adaptive-β DPO بدون نیاز به انتخاب دستی ضریب β، رفتار آموزشی قابل‌کنترل‌تر و پایدارتر نسبت به DPO با β ثابت ارائه می‌دهد، بدون آنکه ساختار اصلی تابع زیان DPO تغییر یابد. لازم به تأکید است که هدف این آزمایش‌ها ارزیابی نهایی کیفیت پاسخ‌ها نبوده، بلکه تحلیل پایداری و رفتار دینامیکی فرآیند آموزش در شرایط کنترل‌شده بوده است. نتایج به‌دست‌آمده نشان می‌دهند که تنظیم تطبیقی β می‌تواند بدون افزایش پیچیدگی محاسباتی، نقش مؤثری در پایدارسازی فرآیند همترازی مبتنی بر ترجیح ایفا کند. در نهایت، باید توجه داشت که این ارزیابی در مقیاس محدود و با هدف اعتبارسنجی پیاده‌سازی و تحلیل رفتار دینامیکی ضریب β انجام شده است. استفاده از مجموعه‌داده‌های بزرگ‌تر، معیارهای کیفی پیشرفته‌تر نظیر win-rate و اندازه‌گیری دقیق‌تر واگرایی KL نسبت به سیاست مرجع می‌تواند در نسخه‌های آتی، ارزیابی جامع‌تری از اثرگذاری روش پیشنهادی ارائه دهد.

ارزیابی و نتایج:

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. تحلیل‌های تجربی نشان دادند که استفاده از β ثابت، به‌ویژه در مقادیر بزرگ‌تر، می‌تواند

منجر به رشد شدید نُرم گرادیان‌ها، افزایش واگرایی نسبت به سیاست مرجع و کاهش پایداری فرآیند آموزش شود، در حالی که بهبود معناداری در یادگیری ترجیحات انسانی ایجاد نمی‌کند. نتایج حاصل از اجرای DPO با مقادیر مختلف β ثابت نشان داد که افزایش β لزوماً به همگرایی بهتر منجر نشده و در بسیاری از موارد، تنها فشار عددی آموزش را افزایش داده است. در مقابل، روش پیشنهادی Adaptive-β DPO با تنظیم پویا و داده‌محور ضریبβ ، توانست شدت اعمال ترجیحات انسانی را متناسب با وضعیت لحظه‌ای مدل کنترل کند. بررسی رفتار آموزشی نشان داد که در این روش، نُرم گرادیان‌ها در طول آموزش در محدوده‌ای پایدار باقی می‌ماند و از نوسانات شدید مشاهده‌شده در DPO با β ثابت جلوگیری می‌شود. همچنین، انحراف مدل از سیاست مرجع به‌صورت کنترل‌شده‌تری مدیریت شده و فرآیند آموزش رفتاری قابل‌پیش‌بینی‌تر و تفسیرپذیرتر از خود نشان داده است. مزیت اصلی روش پیشنهادی آن است که بدون تغییر ساختار تابع زیانDPO ، بدون نیاز به مدل پاداش مجزا و بدون استفاده از یادگیری تقویتی، پایداری آموزش را بهبود می‌دهد. از این منظر، Adaptive-β DPOرا می‌توان به‌عنوان یک تعمیم ساده و مؤثر از DPO استاندارد در نظر گرفت که وابستگی این الگوریتم به تنظیم دستی یک پارامتر بحرانی را کاهش می‌دهد. نتایج این پژوهش نشان می‌دهد که تنظیم تطبیقی شدت ترجیحات انسانی می‌تواند نقش مهمی در بهبود پایداری و کنترل رفتار آموزشی الگوریتم‌های هم‌ترازی مبتنی بر ترجیح ایفا کند و مبنای مناسبی برای توسعه نسخه‌های پایدارتر DPO فراهم آورد.

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.