مروری بر مدل‌های زبانی

سپهر کراچی

۱۴۰۵

فهرست مطالب

۱ آموزش (Training) ۳
۱.۱ داده‌ها (Data) ۳
۱.۱.۱ پیش‌آموزش (Pre-Training) ۳
۲.۱.۱ پس‌آموزش (Post-Training) ۵
۲.۱ روش (Method) ۶
۱.۲.۱ پیش‌آموزش (Pre-Training) ۶
۲.۲.۱ پس‌آموزش (Post-Training) ۶

۲ ساختار (Architecture) ۷
۱.۲ کدگذاری موقعیت (Positional Encoding) ۷
۲.۲ بردارکننده (Tokenizer) ۷
۳.۲ شبکه عصبی (MLP) ۸
۴.۲ تابع فعالساز (Activation Function) ۸
۵.۲ نرمالسازی (Normalization) ۹
۶.۲ مکانیزم توجه (Attention) ۹
۷.۲ ساختار ترنسفورمر (Transformer) ۱۰

۳ بهینه سازی (Optimization) ۱۰
۱.۳ الگوریتم (Algorithm) ۱۰
۲.۳ تابع هزینه (Loss Function) ۱۱

۴ تنظیم هایپرپارامتر (Hyperparameter Tuning) ۱۱

۵ قوانین مقیاس پذیری (Scaling Laws) ۱۲
۱.۵ مقدمه ۱۲
۲.۵ اصلاح و توسعه قوانین مقیاس‌پذیری کلاسیک ۱۲
۳.۵ تاثیر معماری و شکل مدل (Model Shape & Architecture) ۱۲
۴.۵ قوانین مقیاس‌پذیری برای مدل‌های ترکیبی از خبرگان (MoE) ۱۳
۵.۵ نقش داده‌ها: کیفیت، ترکیب و چگالی ۱۳
۶.۵ فراتر از خطای مطلق: معیارهای نوین ارزیابی ۱۴
۷.۵ نتیجه‌گیری ۱۴

۶ ارزیابی (Evaluation) ۱۵

۷ یادگیری تقویتی (Reinforcement Learning) ۱۵
۱.۷ الگوریتم بهینه‌سازی تقریبی خط‌مشی (Proximal Policy Optimization - PPO) ۲۰
۲.۷ الگوریتم بهینه‌سازی نسبی خط‌مشی گروهی (Group Relative Policy Optimization - GRPO) ۲۷
۳.۷ الگوریتم بهینه‌سازی توالی گروهی خط‌مشی (Group Sequence Policy Optimization - GSPO) ۳۲
۴.۷ الگوریتم بهینه‌سازی خط‌مشی با برش مجزا و نمونه‌برداری پویا (Decoupled Clip and Dy-namic Sampling Policy Optimization - DAPO) ۳۹
۵.۷ الگوریتم بهینه‌سازی خط‌مشی تطبیقی نرم (Soft Adaptive Policy Optimization - SAPO) ۴۵
۶.۷ الگوریتم بهینه‌سازی خط‌مشی با ضریب اهمیت برش‌خورده (Clipped IS-weight Policy Optimization - CISPO) ۵۱
۷.۷ الگوریتم یادگیری تقویتی با پاداش‌های اعتبارسنجی‌پذیر (Reinforcement Learning with Verifiable Rewards - RLVR) ۵۷
۸.۷ الگوریتم یادگیری تقویتی با پاداش‌های قابل اعتبارسنجی و بهینه‌سازی گروهی (OlmoRL: RLVR + GRPO) ۶۲

۸ منابع **۷۰**

۱ آموزش (Training)

۱.۱ داده‌ها (Data)

۱.۱.۱ پیش‌آموزش (Pre-Training)

- sd**a**
- sd**a**
- sd**a**
- sd**a**

- **DeepSeek-V4**: پیکره داده‌های مرحله پیش‌آموزش (Pre-Training) این خانواده از مدل‌ها—شامل دو نسخه پایه DeepSeek-V4-Flash و DeepSeek-V4-Pro—با هدف ارتقای بنیادین ظرفیت‌های استدلالی، مهارت‌های کدنویسی، گسترش دانش چندزبانه و پشتیبانی بومی از طول زمینه یک میلیون توکن طراحی و گردآوری شده است [؟]. این پیکره بر بستر تجارب مدل DeepSeek-V3 [؟] استوار بوده و از پالایش‌های الگوریتمی پیشرفته بهره می‌برد.

کلیات و شالوده داده‌ها (Foundational Data Composition) مجموع حجم کل توکن‌های مصرف‌شده در فاز پیش‌آموزش بیش از ۳۲ تریلیون توکن (> 32T) است. مدل DeepSeek-V4-Flash بر روی 32T توکن و مدل بزرگ‌تر DeepSeek-V4-Pro بر روی 33T توکن آموزش دیده‌اند. این پیکره تنوع گسترده‌ای از متون وب پالایش‌شده، کدهای برنامه‌نویسی، ادبیات ریاضیات، اسناد علمی طویل و متون چندفرهنگی را در بر می‌گیرد.

پالایش داده‌های وب و جلوگیری از فروپاشی مدل (Web Data Filtering & Model Collapse) در بخش متون استخراج‌شده از وب، فیلترهای الگوریتمی نوینی برای شناسایی و حذف متون قالبی (Templated) و محتواهای تولیدشده به صورت خودکار توسط سایر مدل‌های هوش مصنوعی اعمال شده است.

- **تشریح پدیده فروپاشی مدل (Model Collapse):** مطابق یافته‌های ژو و همکاران [؟]، آموزش بازگشتی مدل‌ها بر داده‌های سنتتیک و تولیدشده توسط مدل‌های نسل پیشین، منجر به افت واریانس بازنمایی، تحلیل رفتن دم‌های توزیع احتمالاتی (Long-Tail Distribution) و در نهایت زوال تنوع زبانی و استدلالی مدل می‌شود. فیلترینگ سخت‌گیرانه داده‌های وب در DeepSeek-V4 مانع از وقوع این پدیده شده است.

پیکره تخصصی ریاضیات، کدنویسی و داده‌های عاملیتی (Code, Math & Agentic Data) متون ریاضیاتی و مخازن کد هسته اصلی داده‌ها را تشکیل می‌دهند. ویژگی منحصربه‌فرد این نسخه، تزریق **داده‌های عاملیتی (Agentic Data)** در فاز میانی آموزش (Mid-Training) است.

- **داده‌های عاملیتی (Agentic Data):** به توالی‌هایی اطلاق می‌شود که شامل تعاملات گام‌به‌گام با محیط، فراخوانی ابزارها (Tool Calling)، تحلیل خروجی مفسر کد و بازخورد خطاها هستند. این داده‌ها مدل پایه را پیش از ورود به فاز پس‌آموزش، با منطق عامل‌های خودمختار آشنا می‌سازند.
- **فاز میانی آموزش (Mid-Training):** مرحله‌ای پیوسته در اواخر پیش‌آموزش عمومی که در آن توزیع داده‌ها به سمت سناریوهای پیچیده استدلالی و تعاملی سوق داده می‌شود تا فضای پنهان مدل برای یادگیری پس‌آموزش تقویت گردد.

داده‌های چندزبانه و دانش دم‌بلند (Multilingual & Long-Tail Knowledge) به منظور بهبود بازنمایی مفاهیم در زبان‌های مختلف و ارتقای فهم ظرایف فرهنگی، سهم داده‌های چندزبانه به طور چشمگیری افزایش یافته است. این راهبرد یادگیری حقایق نادر و دانش دم‌بلند (Long-Tail Knowledge)—دانشی که فراوانی تکرار اندکی در اسناد وب عمومی دارد—را بهبود بخشیده است.

گردآوری متون طویل و آکادمیک (Long-Document Curation) با توجه به هدف‌گذاری پنجره زمینه یک میلیونی (1M)، بخش ویژه‌ای از خط لوله داده به متون طولانی با پیوستگی معنایی بالا نظیر مقالات علمی، گزارش‌های فنی، کتاب‌ها و تحلیل‌های پژوهشی ساختاریافته اختصاص یافته است تا ساختار وابستگی‌های دوربرد در مدل پایدار شود.

پیش‌پردازش، توکن‌سازی و راهبردهای قالب‌بندی (Pre-processing & Formatting) فرآیند آماده‌سازی توکن‌ها با تکنیک‌های زیر اجرا گردیده است:

- **فضای واژگان (Vocabulary):** حفظ ساختار توکنایزر DeepSeek-V3 با اندازه واژگان ۱۲۸ هزار توکن (128K) و معرفی توکن‌های کنترلی جدید جهت تعیین مرزهای بافتار (Context Construction).
- **پر کردن بخش میانی (Fill-in-the-Middle - FIM):** استفاده از سازوکار تقطیع و بازآرایی متن به قالب پیشوند، پسوند و میانه برای ایجاد قابلیت درک دوطرفه در متن و تکمیل کدها بر اساس رویکرد [؟].
- **بسته‌بندی اسناد (Document Packing):** با الهام از راهبرد دینگ و همکاران [؟]، جهت جلوگیری از هدررفت محاسباتی ناشی از پدگذاری (Padding) و جلوگیری از آسیب ناشی از برش ناگهانی انتهای متون (Sample Truncation)، اسناد کوتاه متعدد درون یک پنجره زمینه واحد ادغام می‌شوند.
- **ماسک‌گذاری توجه در سطح نمونه (Sample-Level Attention Masking):** برخلاف رویکرد آموزش پیوسته پیشین، توجه بین‌نمونه‌ای در یک توالی ادغام‌شده به طور کامل توسط ماسک مسدود می‌شود تا از نشت بافتار (Context Contamination) بین اسناد غیرمرتبط جلوگیری گردد.

برنامه زمانی طول زمینه و اندازه دسته (Curriculum & Scheduling) فرآیند پیش‌آموزش داده‌ها دارای یک برنامه زمانی فزاینده (Curriculum) در دو بعد طول زمینه و اندازه دسته بوده است:

- **برنامه طول دنباله:** آموزش مدل‌ها ابتدا با طول دنباله 4K آغاز شده و به تدریج به 16K، 64K و نهایتاً به 1M توکن ارتقا یافته است:

$$L \in \{4\text{K} \rightarrow 16\text{K} \rightarrow 64\text{K} \rightarrow 1\text{M}\} \tag{۱}$$

- **زمان‌بندی اندازه دسته:** اندازه دسته توکن‌ها از مقادیر اندک آغاز و تا سطوح پایدار مقیاس‌بندی شده است؛ برای نسخه Flash تا سقف 75.5M توکن و برای نسخه Pro تا سقف 94.4M توکن در هر گام بهینه‌سازی.
- **گذار توجه متراکم به کم‌پشت:** در طول ۱ تریلیون توکن نخست (1T)، مدل با سازوکار توجه متراکم (Dense Attention) گرم‌سازی شده و از طول زمینه 64K به بعد، سازوکار توجه کم‌پشت (Sparse Attention) فعال و جایگزین شده است.

تحلیل‌های آماری و فرمولاسیون ریاضی داده‌ها (Statistical & Mathematical Analysis) حجم توکن کل پردازش‌شده ($\mathcal{D}$) در کل مراحل بهینه‌سازی گام $t \in \{1, \ldots, T\}$ از رابطه زیر پیروی می‌کند:

$$\mathcal{D} = \sum_{t=1}^{T} B(t) \tag{۲}$$

که در آن برای مدل‌های مذکور مقادیر نهایی برابرند با:

$$\mathcal{D}_{\text{Flash}} = 3.2 \times 10^{13}\text{tokens}\ , \quad \mathcal{D}_{\text{Pro}} = 3.3 \times 10^{13}\text{tokens} \tag{۳}$$

در فرآیند بسته‌بندی اسناد با اعمال ماسک‌گذاری سطحی، ماتریس ماسک علّی-نمونه‌ای $M \in \mathbb{R}^{L_{\text{seq}} \times L_{\text{seq}}}$ بر روی شباهت‌های لایه توجه به صورت زیر تعریف می‌شود:

$$M_{i,j} = \begin{cases} 0 & \text{if} j \leq i \text{ and } \text{DocID}(i) = \text{DocID}(j) \\ -\infty & \text{otherwise} \end{cases} \tag{۴}$$

که در آن $\text{DocID}(i)$ شناسه سندی است که توکن i-ام به آن تعلق دارد. در نتیجه، ماتریس توجه همواره ارتباط متقابل اسناد ادغام‌شده را به صفر مطلق می‌رساند:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}} + M\right)V \tag{۵}$$

همچنین در زیرساخت موازی‌سازی بافتار (Context Parallelism) همراه با فشرده‌سازی توجه، به منظور پیشگیری از هدررفت توکن‌ها در مرزهای توالی به واسطه ضریب فشرده‌سازی لایه‌ها ($m = 4$ در CSA و $m' = 128$ در HCA)، داده‌های ورودی بر مبنای ک.م.م این ضرایب هم‌تراز می‌گردند:

$$L_{\text{block}} = k \cdot \text{lcm}(m, m') = k \cdot \text{lcm}(4, 128) = 128k, \quad k \in \mathbb{N} \tag{۶}$$

این هم‌ترازی تضمین می‌نماید که اتلاف توکن ناشی از باقیمانده طول دنباله ($l_{\text{seq}} \pmod m$) به حداقل ممکن کاهش یابد.

- sda
- sda
- sda
- sda
- sda
- sda
- sda
- sda

۲.۱.۱ پس‌آموزش (Post-Training)

- sda
- sda
- sda
- sda
- sda
- sda
- sda

- sda
- sda
- sda
- sda
- sda
- sda

۲.۱ روش (Method)

۱.۲.۱ پیش‌آموزش (Pre-Training)

- sda
- sda
- sda
- sda
- sda
- sda
- sda
- sda
- sda
- sda
- sda
- sda

۲.۲.۱ پس‌آموزش (Post-Training)

- sda
- sda
- sda
- sda
- sda
- sda
- sda

- sda
- sda
- sda
- sda
- sda
- sda

۲ ساختار (Architecture)

۱.۲ کدگذاری موقعیت (Positional Encoding)

- sda
- sda
- sda
- sda
- sda
- sda
- sda
- sda
- sda
- sda
- sda
- sda

۲.۲ بردارکننده (Tokenizer)

- sda
- sda
- sda
- sda
- sda
- sda

- sda
- sda
- sda
- sda
- sda
- sda
- sda

۳.۲ شبکه عصبی (MLP)

- sda
- sda
- sda
- sda
- sda
- sda
- sda
- sda
- sda
- sda
- sda
- sda

۴.۲ تابع فعالساز (Activation Function)

- sda
- sda
- sda
- sda
- sda
- sda
- sda

- sda
- sda
- sda
- sda
- sda
- sda

۵.۲ نرمالسازی (Normalization)

- sda
- sda
- sda
- sda
- sda
- sda
- sda
- sda
- sda
- sda
- sda
- sda

۶.۲ مکانیزم توجه (Attention)

- sda
- sda
- sda
- sda
- sda
- sda
- sda
- sda

- sda
- sda
- sda
- sda
- sda

۷.۲ ساختار ترنسفورمر (Transformer)

- sda
- sda
- sda
- sda
- sda
- sda
- sda
- sda
- sda
- sda
- sda
- sda

۳ بهینه سازی (Optimization)

۱.۳ الگوریتم (Algorithm)

- sda
- sda
- sda
- sda
- sda
- sda
- sda

- sda
- sda
- sda
- sda
- sda
- sda

۲.۳ تابع هزینه (Loss Function)

- sda
- sda
- sda
- sda
- sda
- sda
- sda
- sda
- sda
- sda
- sda
- sda

۴ تنظیم هایپرپارامتر (Hyperparameter Tuning)

- sda
- sda
- sda
- sda
- sda
- sda
- sda

- sda
- sda
- sda
- sda
- sda
- sda

۵ قوانین مقیاس پذیری (Scaling Laws)

۱.۵ مقدمه

قوانین مقیاس‌پذیری (Scaling Laws) به عنوان یکی از ارکان اصلی در توسعه و درک مدل‌های زبانی بزرگ (LLMs) شناخته می‌شوند. این قوانین که در ابتدا توسط محققانی چون Kaplan و Hoffmann (قانون Chinchilla) مطرح شدند، نشان می‌دهند که عملکرد مدل‌ها با افزایش پارامترها، حجم داده‌ها و بودجه محاسباتی به صورت قابل پیش‌بینی بهبود می‌یابد. با این حال، تحقیقات اخیر نشان داده‌اند که قوانین کلاسیک دارای محدودیت‌هایی هستند؛ از جمله نادیده گرفتن هزینه‌های استنتاج (Inference)، تاثیر شکل معماری (Architecture Shape)، کیفیت داده‌ها و ساختارهای نوینی مانند «ترکیبی از خبرگان» (Mixture-of-Experts یا MoE). این بخش، با ترکیب و تحلیل ۱۹ مقاله علمی جدید، به بررسی ابعاد پنهان و توسعه‌یافته‌ی قوانین مقیاس‌پذیری می‌پردازد.

۲.۵ اصلاح و توسعه قوانین مقیاس‌پذیری کلاسیک

قوانین اولیه مانند Chinchilla فرض می‌کردند که نسبت بهینه داده به پارامتر (D/N) یک مقدار ثابت (حدود ۲۰) است. اما لی و همکاران [۱] با معرفی قانون Farseer نشان می‌دهند که این نسبت ثابت نیست و با افزایش بودجه محاسباتی، نسبت بهینه D/N نیز افزایش می‌یابد. روش Farseer با استفاده از برازش قطعه‌ای دیفرانسیلی (Differential Piecewise Fitting)، خطای برون‌یابی (Extrapolation Error) را نسبت به قانون Chinchilla تا ۴۳۳ درصد کاهش می‌دهد.

از سوی دیگر، چن و همکاران [۲] پدیده Sub-scaling (کاهش نرخ بهبود عملکرد) را بررسی می‌کنند. این پدیده زمانی رخ می‌دهد که مدل‌ها بیش از حد آموزش می‌بینند (Over-training) یا چگالی داده‌ها (افزونگی اطلاعات) بسیار بالا است. این مقاله یک قانون مقیاس‌پذیری زیر-بهینه (Sub-optimal) ارائه می‌دهد که نشان می‌دهد داده‌های با چگالی بالا منجر به رشد توانی (کاهش بازدهی نهایی) و داده‌های با چگالی پایین منجر به رشد خطی در عملکرد می‌شوند.

همچنین، مک‌لیش و همکاران [۳] با معرفی مجموعه Gemstones (شامل بیش از ۴۰۰۰ چک‌پوینت)، شکنندگی قوانین مقیاس‌پذیری نسبت به انتخاب‌های ابرپارامتری (مانند نرخ یادگیری و شکل مدل) را برجسته می‌کنند. آن‌ها روش Convex Hull را برای برازش قوانین پیشنهاد می‌دهند که نسبت به روش‌های پیشین بسیار مقاوم‌تر است.

۳.۵ تاثیر معماری و شکل مدل (Model Shape & Architecture)

یکی از مهم‌ترین نقایص قوانین گذشته، نادیده گرفتن تاثیر شکل مدل (نسبت عمق به عرض) و هزینه‌های استنتاج است.

- **عرض در برابر عمق:** بیان و همکاران [۴] نشان می‌دهند که معماری مدل مستقیماً بر تاخیر استنتاج (Inference Latency) تاثیر می‌گذارد. مدل‌های عریض‌تر و کم‌عمق‌تر (Wider and Shallower) با بودجه

پارامتری یکسان، توان عملیاتی (Throughput) بسیار بالاتری در استنتاج دارند. در همین راستا، لیو و همکاران [۵] پدیده Inverse Depth Scaling را معرفی می‌کنند؛ جایی که خطا با نسبت معکوس عمق ($1/\ell$) کاهش می‌یابد. آن‌ها دلیل این امر را میانگین‌گیری گروهی (Ensemble Averaging) لایه‌ها می‌دانند، به این معنا که لایه‌های متعدد در LLMها کارهای مشابهی انجام می‌دهند و صرفاً با میانگین‌گیری، خطا را کاهش می‌دهند که نشان‌دهنده استفاده ناکارآمد از عمق است.

- **تعامل عصبی (Neural Interaction):** سان و همکاران [۶] با استفاده از مفهوم Average Gradient Outer Product (AGOP)، قانون «تعامل عصبی» را مطرح می‌کنند. آن‌ها نشان می‌دهند که تعمیم‌پذیری خوب نیازمند برهم‌نهی خوش‌خیم (Benign Superposition) است؛ جایی که مدل با انرژی تعاملی پایین، سهم بالایی از حساسیت تعاملی را به دست می‌آورد. نسبت عمق به عرض ($R_{D/W}$) ابزاری برای قرار دادن مدل در این بازه بهینه است.

- **قوانین مقیاس‌پذیری طیفی (Spectral Scaling Laws):** جها و ریگن [۷] نشان می‌دهند که افزایش عرض شبکه‌های پیش‌خور (FFN) عمدتاً ظرفیت‌های کم‌انرژی (دم طیف) را گسترش می‌دهد، در حالی که زیرفضاهای غالب به سرعت اشباع می‌شوند. این قانون مقیاس‌پذیری طیفی نامتقارن نشان می‌دهد که انتخاب عرض FFN یک مبادله بین ظرفیت دم و ظرفیت حالت غالب است.

- **معماری‌های جایگزین (xLSTM و Mamba):** بک و همکاران [۸] قوانین مقیاس‌پذیری معماری xLSTM (با پیچیدگی زمانی خطی) را با ترانسفورمرها مقایسه می‌کنند. نتایج نشان می‌دهد که xLSTM در مبادله محاسبات-خطا (Compute-Loss Trade-off) بر ترانسفورمرها برتری پارتو (Pareto-dominate) دارد. همچنین، در مقاله‌ای دیگر [۹] با رویکرد تفسیرپذیری مکانیسمی، نشان داده شده است که معماری Mamba بازیابی اطلاعات (Associative Recall) را از طریق یادگیری ضمنی توابع هش خطی انجام می‌دهد و کران‌های بالایی برای اندازه واژگان و ابعاد حالت بر اساس لم Johnson-Lindenstrauss ارائه می‌دهد.

۴.۵ قوانین مقیاس‌پذیری برای مدل‌های ترکیبی از خبرگان (MoE)

مدل‌های MoE به دلیل جداسازی تعداد کل پارامترها از هزینه محاسباتی (FLOPs)، نیازمند قوانین مقیاس‌پذیری مختص به خود هستند.

- **اهرم کارایی (Efficiency Leverage):** تیان و همکاران [۱۰] معیار EL را برای سنجش برتری محاسباتی MoE نسبت به مدل‌های متراکم (Dense) معرفی می‌کنند. آن‌ها نشان می‌دهند که نسبت فعال‌سازی (Activation Ratio) محرک اصلی کارایی است و دانه‌بندی خبرگان (Expert Granularity) به عنوان یک تعدیل‌کننده غیرخطی عمل می‌کند.

- **مبادله پارامتر و FLOPs:** آبنار و همکاران [۱۱] نشان می‌دهند که برای یک بودجه محاسباتی ثابت، افزایش سطح پراکندگی (Sparsity) — یعنی داشتن پارامترهای کل بیشتر اما پارامترهای فعال کمتر — منجر به کاهش خطای آموزش می‌شود.

- **کارایی حافظه:** برخلاف تصور رایج، لودزیجوسکی و همکاران [۱۲] ثابت می‌کنند که مدل‌های MoE می‌توانند از نظر حافظه نیز کارآمدتر از مدل‌های متراکم باشند، به شرطی که روی توکن‌های بیشتری آموزش ببینند.

- **بازیافت مدل‌ها (Upcycling):** لیو و همکاران [۱۳] به بررسی تبدیل مدل‌های متراکم از پیش‌آموزش‌دیده به MoE می‌پردازند. آن‌ها یک قانون مقیاس‌پذیری مشترک کشف کردند که نشان می‌دهد Upcycling تنها زمانی کارآمدتر از آموزش از صفر (From-scratch) است که هزینه اولیه (Sunk Cost - توکن‌های آموزش مدل متراکم) و اندازه مدل نسبتاً کوچک باشند.

۵.۵ نقش داده‌ها: کیفیت، ترکیب و چگالی

داده‌ها سوخت اصلی قوانین مقیاس‌پذیری هستند و کیفیت آن‌ها به اندازه کمیتشان اهمیت دارد.

- **کیفیت داده‌ها:** سوبرامانیام و همکاران [۱۴] یک پارامتر بدون بعد کیفیت به نام $Q \in (0, 1]$ را معرفی کرده و قانون Chinchilla را برای در نظر گرفتن کیفیت داده‌ها بسط می‌دهند. آن‌ها نشان می‌دهند که داده‌های با کیفیت بالا می‌توانند نیاز به اندازه مدل و محاسبات را به شدت کاهش دهند و اندازه موثر نمونه به صورت $D_{eff} = D \cdot Q^{\gamma}$ مقیاس می‌یابد.

- **ترکیب بهینه داده‌ها (Data Mixtures):** انتخاب نسبت ترکیب دامنه‌های مختلف داده معمولاً با آزمون و خطا انجام می‌شود. شوکور و همکاران [۱۵] روشی سیستماتیک ارائه می‌دهند که با استفاده از برازش قوانین مقیاس‌پذیری روی اجراهای مقیاس کوچک، می‌توان وزن‌های بهینه دامنه را برای آموزش مدل‌های بسیار بزرگ پیش‌بینی کرد.

- **ساختار سلسله‌مراتبی داده‌ها:** کاگنتا و همکاران [۱۶] با استفاده از مدل سلسله‌مراتب تصادفی (RHM)، نشان می‌دهند که شبکه‌های عصبی پیچیدگی‌های داده را لایه به لایه یاد می‌گیرند. آن‌ها ثابت می‌کنند که شبکه‌های پیچشی (CNNs) به دلیل اشتراک وزن و اتصال محلی، در یادگیری داده‌های سلسله‌مراتبی سریع‌تر از ترانسفورمرها مقیاس می‌یابند.

۶.۵ فراتر از خطای مطلق: معیارهای نوین ارزیابی

ارزیابی مدل‌ها صرفاً بر اساس خطای آنتروپی متقاطع (Cross-Entropy) تصویر کاملی از توانایی‌های مدل ارائه نمی‌دهد.

- **قوانین مقیاس‌پذیری نسبی (Relative Scaling Laws):** هلد و همکاران [۱۷] استدلال می‌کنند که ارزیابی‌های تجمیعی، تفاوت عملکرد در زیرگروه‌ها را پنهان می‌کند. آن‌ها قوانین نسبی را معرفی می‌کنند تا نشان دهند چگونه شکاف عملکرد بین توزیع‌ها با افزایش مقیاس تغییر می‌کند. نتایج نشان می‌دهد که مقیاس‌پذیری یک تساوی‌بخش جهانی نیست؛ در حالی که دامنه‌های علمی در MMLU به همگرایی می‌رسند، خطرات هوش مصنوعی (مانند توانایی فریبکاری) با افزایش مقیاس واگرا شده و افزایش می‌یابند.

- **احتمال مبتنی بر رتبه (Relative-Based Probability):** یو و همکاران [۱۸] معیار RBP را معرفی می‌کنند که احتمال قرار گرفتن توکن صحیح در میان k پیش‌بینی برتر را می‌سنجد. آن‌ها نشان می‌دهند که $-\log(RBP_k)$ نیز از یک قانون توانی پیروی می‌کند. این دیدگاه پدیده ظهور توانایی‌ها (Emergence) را نه به عنوان یک جهش ناگهانی، بلکه به عنوان یک اثر ماکروسکوپی قابل پیش‌بینی از مقیاس‌پذیری میکروسکوپی هموار توضیح می‌دهد.

- **قانون چگال‌سازی (Densing Law):** در نهایت، شیائو و همکاران [۱۹] برای ارزیابی روند پیشرفت کارایی LLMها در طول زمان، مفهوم چگالی قابلیت (Capability Density) را معرفی می‌کنند. مشاهدات تجربی آن‌ها نشان می‌دهد که حداکثر چگالی قابلیت مدل‌های متن‌باز تقریباً هر ۵.۳ ماه دو برابر می‌شود که نشان‌دهنده کاهش نمایی هزینه‌های استنتاج و نیازهای پارامتری برای رسیدن به یک سطح عملکرد ثابت است.

۷.۵ نتیجه‌گیری

بررسی جامع این مقالات نشان‌دهنده یک تغییر پارادایم در درک ما از قوانین مقیاس‌پذیری است. ما از دوران «بزرگتر بودن همیشه بهتر است» (صرفاً با افزایش پارامتر و داده) عبور کرده‌ایم. قوانین مقیاس‌پذیری نوین اکنون چندوجهی شده‌اند و باید هزینه‌های استنتاج، شکل معماری (عمق/عرض)، کیفیت و چگالی داده‌ها، ساختارهای پراکنده (MoE) و معیارهای ارزیابی نسبی را در بر بگیرند. این بینش‌های جدید به محققان و مهندسان اجازه می‌دهد تا مدل‌های زبانی آینده را نه تنها قدرتمندتر، بلکه از نظر محاسباتی، حافظه و استنتاج به مراتب کارآمدتر و هدفمندتر طراحی کنند.

۶ ارزیابی (Evaluation)

- sda
- sda
- sda
- sda
- sda
- sda
- sda
- sda
- sda
- sda
- sda
- sda
- sda

۷ یادگیری تقویتی (Reinforcement Learning)

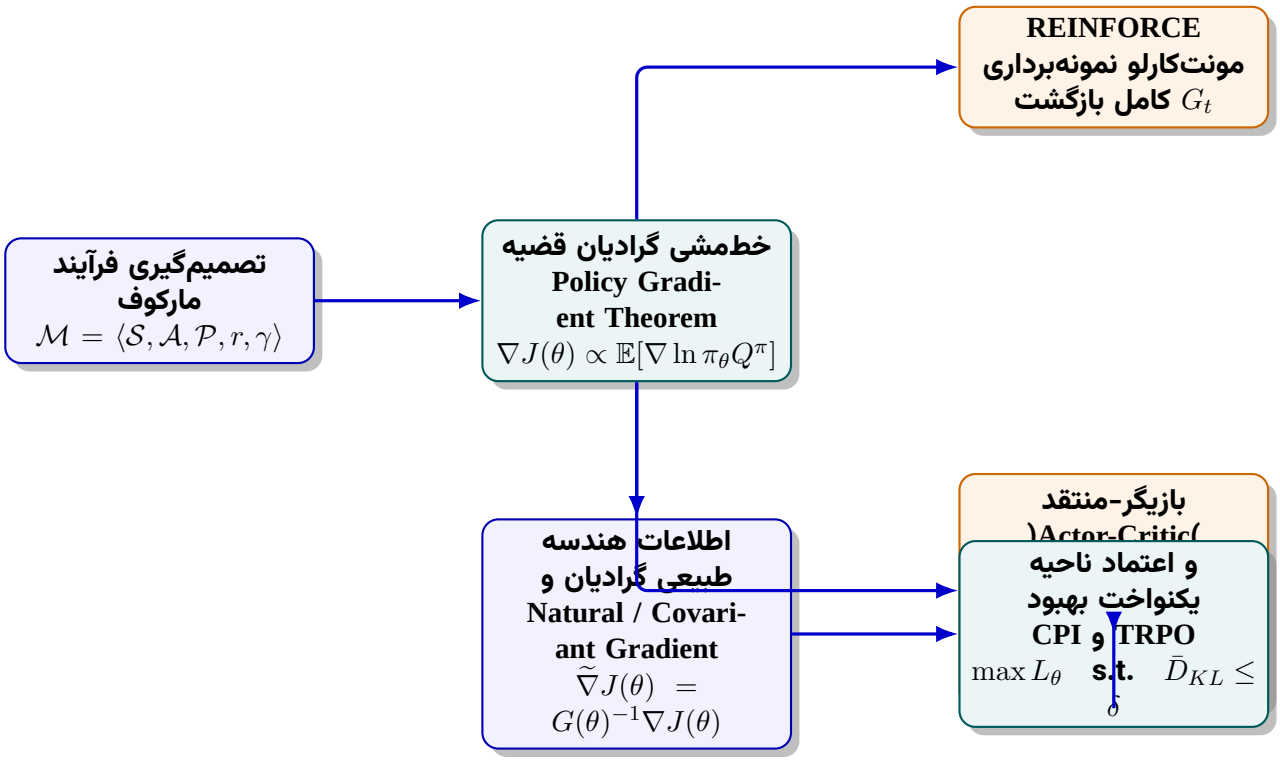
۰ مبانی و تئوری الگوریتم‌های مبتنی بر خط‌مشی (Policy-Based RL Theory)

مقدمه و انگیزه‌های بنیادین

روش‌های یادگیری تقویتی سنتی مبتنی بر ارزش (Value-Based)، نظیر Q-Learning، خط‌مشی را از طریق بیشینه‌سازی مقادیر عمل ($a = \text{argmax}_a Q(s,a)$) استخراج می‌کنند. این رویکردها با چالش‌های نظری و محاسباتی متعددی روبه‌رو هستند [۲۰]:

- **انقطاع در خط‌مشی (Discontinuity):** تغییرات بسیار کوچک در مقادیر تقریب‌زده‌شده $Q(s,a)$ می‌تواند منجر به تغییرات ناگهانی و ناپیوسته در خط‌مشی انتخابی شود که مانع از همگرایی پایدار در فضاهای پارامتری پیوسته می‌گردد [۲۱].
- **فضاهای عمل پیوسته (Continuous Action Spaces):** حل مسئله بیشینه‌سازی پیوسته $\max_{a\in\mathcal{A}} Q(s,a)$ در هر گام محاسباتی، به‌ویژه با تقریب‌زننده‌های عصبی، بسیار پرهزینه است.
- **خط‌مشی‌های تصادفی بهینه (Stochastic Policies):** در محیط‌های با مشاهدات جزئی (POMDPs) و در شرایط تقریب تابع، خط‌مشی بهینه ممکن است ماهیتاً تصادفی باشد (مانند بازی‌های با اطلاعات ناقص یا راهروی کوتاه با اعمال معکوس [۲۰]).

الگوریتم‌های مبتنی بر خط‌مشی (Policy-Based) با پارامترسازی مستقیم خط‌مشی $\pi_\theta(a|s)$ توسط بردار پارامتر $\theta \in \mathbb{R}^d$، تابع هدف عملکردی $J(\theta)$ را به صورت هموار و مستقیم بهینه‌سازی می‌کنند.



قضیه گرادیان خط‌مشی (Policy Gradient Theorem)

فرآیند تصمیم‌گیری مارکوف را با توزیع اولیه ρ_0، پاداش کران‌دار $r(s,a)$ و ضریب تنزیل $\gamma \in [0,1)$ در نظر می‌گیریم. تابع ارزش حالت $V^\pi(s)$ و ارزش حالت-عمل $Q^\pi(s,a)$ به صورت زیر تعریف می‌شوند:

$$V^\pi(s) \triangleq \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, s_0 = s \right], \tag{۷}$$

$$Q^\pi(s,a) \triangleq r(s,a) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s,a) V^\pi(s'). \tag{۸}$$

همچنین تابع مزیت (Advantage Function) برابر است با $A^\pi(s,a) \triangleq Q^\pi(s,a) - V^\pi(s)$. معیار عملکرد هدف برای حالت اپیزودیک برابر با $J(\theta) \triangleq V^{\pi_\theta}(s_0)$ است.

اثبات تحلیلی در حالت اپیزودیک: با مشتق‌گیری مستقیم نسبت به θ (برای سادگی نماد، $\pi \equiv \pi_\theta$):

$$\begin{aligned} \nabla V^\pi(s) &= \nabla \left[\sum_{a \in \mathcal{A}} \pi(a|s) Q^\pi(s,a) \right] \\ &= \sum_{a \in \mathcal{A}} \left[\nabla \pi(a|s) Q^\pi(s,a) + \pi(a|s) \nabla Q^\pi(s,a) \right]. \end{aligned} \tag{۹}$$

با توجه به اینکه پاداش و ماتریس انتقال محیط مستقل از θ هستند:

$$\nabla Q^\pi(s,a) = \nabla \left[r(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) V^\pi(s') \right] = \gamma \sum_{s'} \mathcal{P}(s'|s,a) \nabla V^\pi(s'). \tag{۱۰}$$

با جایگذاری در رابطه قبل و گشودن رابطه به صورت بازگشتی روی زمان (Unrolling):

$$\begin{aligned}\nabla V^{\pi}(s) &= \sum_{a} \nabla \pi(a|s) Q^{\pi}(s,a) + \gamma \sum_{s'} P_{\pi}(s'|s) \nabla V^{\pi}(s') \\ &= \sum_{x \in \mathcal{S}} \sum_{t=0}^{\infty} \gamma^{t} \Pr(s \to x, t, \pi) \sum_{a} \nabla \pi(a|x) Q^{\pi}(x,a). \end{aligned} \tag{۱۱}$$

با تعریف توزیع اشغال حالات تنزیل‌شده $\rho_{\pi}(s) \triangleq \sum_{t=0}^{\infty} \gamma^{t} \Pr(s_0 \to s, t, \pi)$، گرادیان عملکرد به دست می‌آید [۲۲]:

$$\nabla J(\theta) = \sum_{s \in \mathcal{S}} \rho_{\pi}(s) \sum_{a \in \mathcal{A}} \nabla \pi_{\theta}(a|s) Q^{\pi}(s,a) = \mathbb{E}_{s \sim \rho_{\pi}, a \sim \pi_{\theta}} \left[\nabla_{\theta} \ln \pi_{\theta}(a|s) Q^{\pi}(s,a) \right]. \tag{۱۲}$$

حالت نامتناهی با پاداش میانگین (Continuing / Average-Reward Case): برای مسائل پیوسته، عملکرد با نرخ پاداش میانگین $r(\pi) \triangleq \lim_{h \to \infty} \frac{1}{h} \sum_{t=1}^{h} \mathbb{E}[R_t]$ تعریف می‌شود. با تعریف تابع ارزش دیفرانسیلی $V^{\pi}(s) = \sum_{a} \pi(a|s)[r(s,a) - r(\pi) + \sum_{s'} \mathcal{P}(s'|s,a) V^{\pi}(s')]$ و ضرب در توزیع ایستا $\mu(s)$، رابطه زیر با خنثی شدن تغییرات ارزش حاصل می‌شود [۲۰]:

$$\nabla J(\theta) = \nabla r(\pi) = \sum_{s \in \mathcal{S}} \mu(s) \sum_{a \in \mathcal{A}} \nabla \pi_{\theta}(a|s) Q^{\pi}(s,a). \tag{۱۳}$$

الگوریتم REINFORCE و خط مبنا (Baseline)

الگوریتم REINFORCE [۲۳] از بازگشت تجربی مسیر $G_t = \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$ به عنوان برآوردگر ناریب $Q^{\pi}(S_t, A_t)$ استفاده می‌کند:

$$\theta_{t+1} = \theta_t + \alpha \gamma^{t} G_t \nabla_{\theta} \ln \pi_{\theta}(A_t|S_t). \tag{۱۴}$$

برای کاهش واریانس، یک تابع خط مبنا $b(s)$ مستقل از عمل تفریق می‌گردد. از آنجا که:

$$\sum_{a \in \mathcal{A}} \nabla \pi_{\theta}(a|s) b(s) = b(s) \nabla_{\theta} \sum_{a \in \mathcal{A}} \pi_{\theta}(a|s) = b(s) \nabla_{\theta}(1) = \mathbf{0}, \tag{۱۵}$$

تفریق خط مبنا هیچ اریبی در تخمین گرادیان ایجاد نکرده و واریانس را به شدت کاهش می‌دهد. انتخاب $b(s) = V(s)$ ضریب گرادیان را به مزیت $A(s,a)$ تبدیل می‌کند [۲۰].

هندسه اطلاعات و گرادیان خط‌مشی هم‌وردا (Covariant Policy Search)

گرادیان استاندارد اقلیدسی به شدت وابسته به نحوه پارامترسازی است (Non-Covariant). برای تحقق استقلال از پارامترسازی، بگنل و اشنایدر [۲۴] منیفلد آماری توزیع کل مسیرها (Path Space) را تعریف کردند:

$$P(\xi; \theta) = \rho_0(s_0) \prod_{t=0}^{T-1} \pi_{\theta}(a_t|s_t) \mathcal{P}(s_{t+1}|s_t, a_t). \tag{۱۶}$$

طبق **قضیه چنتسوف (Chentsov's Theorem)**، تنها متریک ریمانی که نسبت به تبدیل‌های آماری و تعویض متغیرها ناوردا است، **متریک فیشر-رائو (Fisher-Rao Metric)** است [۲۴، ۲۵]:

$$G(\theta) = \mathbb{E}_{\xi \sim P(\cdot;\theta)} \left[\nabla_{\theta} \ln P(\xi;\theta) \nabla_{\theta} \ln P(\xi;\theta)^{\top} \right] = \mathbb{E}_{s \sim d_{\rho_0,\pi}} \left[\sum_{a \in \mathcal{A}} \pi_{\theta}(a|s) \nabla_{\theta} \ln \pi_{\theta}(a|s) \nabla_{\theta} \ln \pi_{\theta}(a|s)^{\top} \right]. \tag{۱۷}$$

جهت صعود بیشترین شیب هم‌وردا (Natural Policy Gradient) به صورت زیر حاصل می‌شود:

$$\widetilde{\nabla} J(\theta) = G(\theta)^{-1} \nabla J(\theta). \tag{۱۸}$$

کاکاده [۲۵] نشان داد که اگر تابع ارزش با تقریب‌زننده خطی سازگار (Compatible Function Approximation) $f_w(s,a) = w^\top \nabla_\theta \ln \pi_\theta(a|s)$ تقریب زده شود، وزن‌های بهینه کمترین مربعات دقیقاً منطبق بر گرادیان طبیعی هستند: $w^* = \widetilde{\nabla} J(\theta)$.

تکرار خط‌مشی محافظه‌کارانه (CPI) و کران‌های بهبود

کاکاده و لنگفورد [۲۱] اتحاد تفاوت عملکرد (Performance Difference Lemma) را برای هر دو خط‌مشی دلخواه π و $\tilde{\pi}$ اثبات کردند:

$$\eta(\tilde{\pi}) - \eta(\pi) = \sum_{s \in \mathcal{S}} \rho_{\tilde{\pi}}(s) \sum_{a \in \mathcal{A}} \tilde{\pi}(a|s) A^\pi(s,a). \tag{۱۹}$$

با تقریب توزیع حالات توسط خط‌مشی فعلی، تقریب محلی زیر تعریف می‌شود:

$$L_\pi(\tilde{\pi}) \triangleq \eta(\pi) + \sum_{s \in \mathcal{S}} \rho_\pi(s) \sum_{a \in \mathcal{A}} \tilde{\pi}(a|s) A^\pi(s,a). \tag{۲۰}$$

در الگوریتم CPI، به‌روزرسانی خط‌مشی به صورت ترکیب تصادفی $\pi_{\text{new}} = (1-\alpha)\pi + \alpha\pi'$ با سیاست حریصانه π' صورت می‌گیرد. با استفاده از نظریه جفت‌شدگی متغیرهای تصادفی (α-Coupling)، کران بهبود زیر اثبات می‌شود [۲۱]:

$$\eta(\pi_{\text{new}}) \geq L_\pi(\pi_{\text{new}}) - \frac{2\gamma\epsilon}{(1-\gamma)^2}\alpha^2, \quad \text{where} \quad \epsilon = \max_s |\mathbb{E}_{a\sim\pi'}[A^\pi(s,a)]|. \tag{۲۱}$$

بهینه‌سازی در ناحیه اعتماد (TRPO)

شولمن و همکاران [۲۶] کران بهبود CPI را با استفاده از تئوری اختلال اپراتوری به تمامی فضاهای خط‌مشی تصادفی بر حسب واگرایی تغییرات کل (D_{TV}) و واگرایی کولبک-لایبلر (D_{KL}) تعمیم دادند:

$$\eta(\pi_{\text{new}}) \geq L_{\pi_{\text{old}}}(\pi_{\text{new}}) - \frac{4\epsilon\gamma}{(1-\gamma)^2} \max_s D_{TV}(\pi_{\text{old}}(\cdot|s) \parallel \pi_{\text{new}}(\cdot|s))^2. \tag{۲۲}$$

با استفاده از نامساوی پینسکر ($D_{TV}^2 \leq \frac{1}{2} D_{KL}$)، مسئله بهینه‌سازی در ناحیه اعتماد (Trust Region) به فرم زیر حاصل می‌شود:

$$\max_\theta L_{\theta_{\text{old}}}(\theta) \quad \text{to subject} \quad \bar{D}_{KL}^{\rho_{\theta_{\text{old}}}}(\theta_{\text{old}}, \theta) \leq \delta. \tag{۲۳}$$

حل تقریبی این مسئله درجه دوم مقید توسط روش گرادیان مزدوج (Conjugate Gradient) و ضرب ماتریس فیشر-بردار (Fisher-Vector Product) به فرم زیر اجرا می‌گردد [۲۶]:

$$\theta_{\text{new}} = \theta_{\text{old}} + \sqrt{\frac{2\delta}{g^\top A^{-1} g}} A^{-1} g, \quad g = \nabla_\theta L_{\theta_{\text{old}}}(\theta), \quad A = \nabla_\theta^2 \bar{D}_{KL}. \tag{۲۴}$$

Algorithm 1 Trust Region Policy Optimization (TRPO)

Require: Initial policy parameters θ_0, KL step size limit δ, line search backtrack coefficient $\zeta \in (0,1)$

for $k = 0, 1, 2, \ldots$ **do**
  Collect trajectory rollouts $\mathcal{D}_k = \{\tau_i\}$ using policy π_{θ_k}.
  Estimate advantages $\hat{A}^{\pi_{\theta_k}}(s_t, a_t)$ using GAE or state-value baseline.
  Compute objective gradient: $g_k = \nabla_\theta L_{\theta_k}(\theta)|_{\theta=\theta_k}$.
  Compute Fisher Information Matrix-vector product function $v \mapsto A_k v = \nabla_\theta \left((\nabla_\theta \bar{D}_{KL})^\top v\right)$.
  Solve $A_k x_k \approx g_k$ using the Conjugate Gradient algorithm.
  Compute maximal step length: $\beta_k = \sqrt{2\delta/(x_k^\top A_k x_k)}$.
  Update direction: $\Delta\theta_k = \beta_k x_k$.
  Perform backtracking line search: find smallest integer $j \geq 0$ such that:
    $L_{\theta_k}(\theta_k + \zeta^j \Delta\theta_k) > 0$ and $\bar{D}_{KL}(\theta_k, \theta_k + \zeta^j \Delta\theta_k) \leq \delta$.
  Update parameter: $\theta_{k+1} = \theta_k + \zeta^j \Delta\theta_k$.
end for

تئوری فرآیندهای مارکوف منظم‌شده (Regularized MDPs)

گایست، شرر و پیتکین [۲۷] با معرفی عملگرهای بلمن منظم‌شده از طریق تبدیل لژاندر-فنچل (Legendre-Fenchel Conjugate)، تئوری یکپارچه‌ای برای منظم‌سازی آنتروپی و واگرایی برگمن ارائه دادند:

$$T_{\pi,\Omega} V \triangleq T_\pi V - \Omega(\pi) = r_\pi - \Omega(\pi) + \gamma P_\pi V, \tag{۲۵}$$

$$T_{*,\Omega} V \triangleq \max_{\pi \in \Delta(\mathcal{A})} T_{\pi,\Omega} V = \Omega^*(q), \quad \text{where} \quad q(s,a) = r(s,a) + \gamma \mathbb{E}_{s'}[V(s')]. \tag{۲۶}$$

عملگرهای $T_{\pi,\Omega}$ و $T_{*,\Omega}$ هر دو γ-انقباض در نرم بی‌نهایت هستند. در الگوریتم فرود آینه‌ای تکرار خط‌مشی اصلاح‌شده (MD-MPI)، با جایگزینی $\Omega(\pi)$ با واگرایی برگمن $D_\Omega(\pi \parallel \pi_k)$ نسبت به خط‌مشی مرحله قبل:

$$\pi_{k+1} = \mathrm{argmax}_\pi \left(\langle q_k, \pi \rangle - D_\Omega(\pi \parallel \pi_k)\right). \tag{۲۷}$$

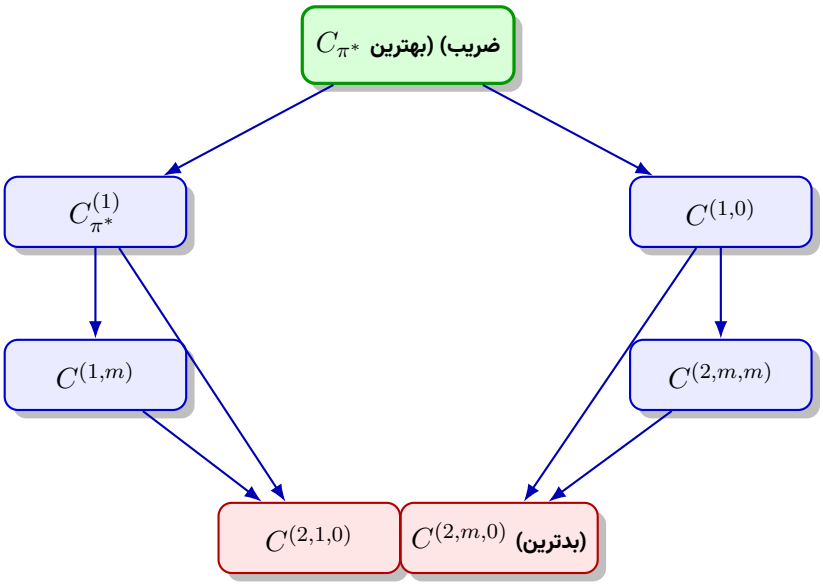
با استفاده از **اتحاد سه‌نقطه‌ای برگمن (Three-Point Identity)**، نرخ همگرایی میانگین حسرت (Regret) به فرم زیر اثبات می‌شود [۲۷]:

$$\frac{1}{K}\sum_{k=1}^{K} \|V^* - V^{\pi_k}\|_\infty \leq \frac{1-\gamma^K}{(1-\gamma)^2}\left[\frac{2\gamma\|V^* - V_0\|_\infty + \max_\pi \|D_\Omega(\pi \parallel \pi_0)\|_\infty}{K}\right] = \mathcal{O}\left(\frac{1}{K}\right). \tag{۲۸}$$

مقایسه ساختاریافته الگوریتم‌های تکرار خط‌مشی تقریبی

شرر [۲۸] چهار خانواده الگوریتم‌های تکرار خط‌مشی تقریبی را از منظر ضرایب تمرکزپذیری (Concentrability Coefficients) مقایسه کرده است:

- $C_{\pi^*} \triangleq \|\frac{d_{\mu,\pi^*}}{\nu}\|_\infty$ (تنها توزیع خط‌مشی بهینه π^* را می‌سنجد و بهترین ضریب است).
- $C^{(1,0)} \triangleq (1-\gamma)\sum_{i=0}^{\infty} \gamma^i \sup_{\pi_1,\ldots,\pi_i} \|\frac{\mu P_{\pi_1} \ldots P_{\pi_i}}{\nu}\|_\infty$.
- $C^{(2,1,0)}$ بدترین ضریب است که خطاهای تمام خط‌مشی‌ها را در بدترین حالت ترکیب می‌کند.



جدول ۱: مقایسه مشخصه‌های نظری الگوریتم‌های تکرار خط‌مشی تقریبی [۲۸]

الگوریتم	کران خطای عملکرد ($\mu(V^* - V^{\pi_k})$)	تعداد تکرارها	پیچیدگی حافظه	نوع خط‌مشی
API	$\frac{C^{(2,1,0)}}{(1-\gamma)^2}\epsilon$	$\mathcal{O}\left(\frac{1}{1-\gamma}\ln\frac{1}{\epsilon}\right)$	$\mathcal{O}(1)$	ایستا و قطعی
CPI	$\frac{C_{\pi^*}}{(1-\gamma)^2}\epsilon$	$\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$	$\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$	ایستا و تصادفی (مخلوط)
PSDP$_\infty$	$\frac{C_{\pi^*}}{(1-\gamma)^2}\epsilon\ln\frac{1}{\epsilon}$	$\mathcal{O}\left(\frac{1}{1-\gamma}\ln\frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\frac{1}{1-\gamma}\ln\frac{1}{\epsilon}\right)$	غیرایستا (حلقه‌ای)
NSPI(m)	$\left(C_{\pi^*}+\frac{\gamma^m C^{(2,m,0)}}{m(1-\gamma^m)}\right)\frac{\epsilon\ln(1/\epsilon)}{(1-\gamma)^2}$	$\mathcal{O}\left(\frac{1}{1-\gamma}\ln\frac{1}{\epsilon}\right)$	$\mathcal{O}(m)$	m-دوره‌ای غیرایستا

همان‌طور که در جدول مشاهده می‌شود، الگوریتم PSDP$_\infty$ بهترین ویژگی‌های هر دو روش API (همگرایی سریع لگاریتمی) و CPI (ضریب تمرکزپذیری بهینه C_{π^*}) را ترکیب می‌کند. الگوریتم NSPI(m) نیز با معرفی حافظه با اندازه ثابت m، خطای ناشی از تقریب را با ضریب هندسی کاهشی γ^m کنترل کرده و تعادل کاملی میان مصرف حافظه و کیفیت عملکرد بر قرار می‌سازد [۲۸، ۲۹].

۱.۷ الگوریتم بهینه‌سازی تقریبی خط‌مشی (Proximal Policy Optimization - PPO)

الگوریتم **PPO** که توسط جان شولمن (John Schulman) و همکارانش در سال ۲۰۱۷ در OpenAI معرفی شد [۳۰]، یکی از برجسته‌ترین، پایدارترین و پرکاربردترین الگوریتم‌های حوزه یادگیری تقویتی عمیق (Deep Reinforcement Learning) به شمار می‌رود. این الگوریتم با حل چالش‌های بنیادین بهینه‌سازی در روش‌های پیشین، تعادلی بهینه میان سادگی پیاده‌سازی، کارایی محاسباتی، بهره‌وری نمونه (Sample Efficiency) و پایداری همگرایی برقرار ساخته است.

تاریخچه، انگیزه و ضرورت پیدایش الگوریتم

پیش از ابداع **PPO**، روش‌های یادگیری تقویتی عمیق به سه دسته عمده با محدودیت‌های ساختاری تقسیم می‌شدند:

- **روش‌های مبتنی بر ارزش (Value-Based Methods):** شاخص‌ترین الگوریتم این دسته، **DQN** [۳۱] است. این روش‌ها گرچه در فضاهای عمل گسسته و محیط بازی‌های آتاری [۳۲] موفق

ظاهر شدند، اما امکان تعمیم کارآمد به فضاهای عمل پیوسته (Continuous Action Spaces) با ابعاد بالا در رباتیک [۳۳، ۳۴] را نداشتند، چرا که یافتن $\arg\max_a Q(s,a)$ در هر گام نیازمند بهینه‌سازی غیرخطی زمان‌بر است. همچنین این روش‌ها از ناپایداری‌های ناشی از تقریب تابع و یادگیری خارج از خط‌مشی رنج می‌برند.

- **روش‌های گرادیان خط‌مشی استاندارد (Vanilla Policy Gradient):** روش‌هایی نظیر REINFORCE [۲۳] و **A2C/A3C** [۳۵] مستقیماً پارامترهای خط‌مشی را بهینه‌سازی می‌کنند. با این حال، به دلیل وابستگی گرادیان به توزیع داده‌های جمع‌آوری‌شده، مجاز به اجرای تنها یک گام گرادیان به ازای هر نمونه داده بودند. انجام چندین گام بهینه‌سازی (Multiple Epochs) روی داده‌های یکسان، منجر به جابجایی شدید خط‌مشی جدید نسبت به خط‌مشی قدیم و در نتیجه سقوط فاجعه‌بار کارایی (Catastrophic Policy Collapse) می‌شد؛ امری که باعث ناکارآمدی شدید در تعداد نمونه‌ها (Sample Inefficiency) بود.
- **روش‌های بهینه‌سازی در ناحیه اطمینان (Trust Region Methods):** کاکاده و لنگفورد [۲۱] و سپس شولمن و همکاران با معرفی الگوریتم **TRPO** [۲۶] نشان دادند که با مقید ساختن تغییرات خط‌مشی از طریق واگرایی کولبک-لایبلر (KL-Divergence)، می‌توان بهبود یکنواخت کارایی (Monotonic Improvement) را تضمین کرد. با این وجود، **TRPO** به محاسبات مرتبه دوم ماتریس اطلاعات فیشر (Fisher Information Matrix)، ماتریس هسین و حل با الگوریتم گرادیان مزدوج (Conjugate Gradient) نیازمند بود. این امر نه تنها پیچیدگی محاسباتی بالایی ایجاد می‌کرد، بلکه با معماری‌های دارای لایه‌های مشترک میان خط‌مشی و ارزش و تکنیک‌های نویزدار مانند Dropout ناسازگار بود.

هدف بنیادین **PPO**، دستیابی به ثبات، قابلیت اطمینان و نرخ همگرایی بالای **TRPO**، صرفاً با تکیه بر بهینه‌سازهای مشتق اول (First-Order Optimizers) نظیر بهینه‌ساز Adam [۳۶] و با سادگی پیاده‌سازی بالا است.

مبانی نظری و اشتقاق‌های کامل ریاضی

یک فرآیند تصمیم‌گیری مارکوف (MDP) با چندتایی متناهی یا نامتناهی $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma, \rho_0)$ مدل‌سازی می‌شود که در آن $\mathcal{S}$ فضای حالات، $\mathcal{A}$ فضای اعمال، $\mathcal{P}(s'|s,a)$ تابع احتمال انتقال، $r(s,a)$ تابع پاداش، $\gamma \in [0,1)$ ضریب تخفیف پاداش و $\rho_0(s_0)$ توزیع حالت اولیه است.

توزیع یک مسیر (Trajectory) $\tau = (s_0, a_0, s_1, a_1, \ldots, s_T)$ تحت خط‌مشی تصادفی پارامتریزه شده $\pi_\theta(a|s)$ برابر است با:

$$P(\tau;\theta) = \rho_0(s_0) \prod_{t=0}^{T-1} \pi_\theta(a_t|s_t)\mathcal{P}(s_{t+1}|s_t,a_t) \tag{۲۹}$$

هدف بهینه‌سازی، بیشینه‌سازی امید ریاضی بازده کل است:

$$J(\theta) = \mathbb{E}_{\tau\sim P(\cdot;\theta)}[R(\tau)] = \int_\tau P(\tau;\theta)R(\tau)\,d\tau \tag{۳۰}$$

که در آن $R(\tau) = \sum_{t=0}^{T} \gamma^t r(s_t,a_t)$ است. با اعمال عملگر گرادیان و استفاده از قاعده مشتق لگاریتمی ($\nabla_\theta P = P\nabla_\theta \log P$):

$$\begin{aligned}\nabla_\theta J(\theta) &= \nabla_\theta \int P(\tau;\theta)R(\tau)\,d\tau = \int \nabla_\theta P(\tau;\theta)R(\tau)\,d\tau \\ &= \int P(\tau;\theta)\nabla_\theta \log P(\tau;\theta)R(\tau)\,d\tau = \mathbb{E}_{\tau\sim\pi_\theta}\left[\nabla_\theta \log P(\tau;\theta)R(\tau)\right]\end{aligned} \tag{۳۱}$$

از آنجا که جملات انتقال وضعیت محیط $\mathcal{P}$ به پارامتر θ وابسته نیستند، داریم $\nabla_\theta \log P(\tau;\theta) = \sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t)$. با بهره‌گیری از اصل علیت و کسر خط مبنای مستقل از عمل $b(s_t) = V(s_t)$ جهت کاهش واریانس، قضیه گرادیان خط‌مشی حاصل می‌شود:

$$\hat{g} = \hat{\mathbb{E}}_t \left[\nabla_\theta \log \pi_\theta(a_t|s_t) \hat{A}_t \right] \tag{۳۲}$$

که در آن $\hat{A}_t = Q(s_t, a_t) - V(s_t)$ برآوردگر تابع مزیت (Advantage Function) است.

با استفاده از مفهوم نمونه‌برداری ترجیحی (Importance Sampling)، هدف خط‌مشی برای داده‌های جمع‌آوری‌شده توسط $\pi_{\theta_{\text{old}}}$ به صورت تابع هدف جایگزین تکرار خط‌مشی محافظه‌کارانه (CPI) [۲۱] تعریف می‌شود:

$$L^{CPI}(\theta) = \hat{\mathbb{E}}_t \left[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)} \hat{A}_t \right] = \hat{\mathbb{E}}_t \left[r_t(\theta) \hat{A}_t \right] \tag{۳۳}$$

که در آن نسبت احتمالات $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$ بوده و $r_t(\theta_{\text{old}}) = 1$ است.

توابع هدف در الگوریتم PPO

الگوریتم **PPO** دو فرمول‌بندی برای محدود ساختن تغییرات گام خط‌مشی ارائه می‌دهد:

۱. تابع هدف با برش جایگزین (Clipped Surrogate Objective): رویکرد اصلی و پیشنهادی مقاله، استفاده از تابع هدف زیر است:

$$L^{\text{CLIP}}(\theta) = \hat{\mathbb{E}}_t \left[\min \left(r_t(\theta) \hat{A}_t,\ \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon) \hat{A}_t \right) \right] \tag{۳۴}$$

که در آن ϵ یک ابرپارامتر (معمولاً $\epsilon = 0.2$) است. این تابع هدف یک کران پایین بدبینانه (Pessimistic Lower Bound) را تشکیل می‌دهد:

- **در شرایط مزیت مثبت ($\hat{A}_t > 0$):** اگر عمل انجام‌شده بهتر از حد انتظار باشد، افزایش احتمال آن مفید است؛ اما به محض اینکه نسبت احتمال از $1+\epsilon$ فراتر رود، گرادیان صفر شده و از تغییرات مفرط جلوگیری می‌کند.
- **در شرایط مزیت منفی ($\hat{A}_t < 0$):** اگر عمل بدتر از حد انتظار باشد، احتمال آن باید کاهش یابد؛ اگر نسبت احتمال از $1-\epsilon$ کمتر شود، مقدار بریده شده و گرادیان صفر می‌شود. اما اگر خط‌مشی اشتباهاً احتمال این عمل بد را افزایش دهد ($r_t \gg 1$)، مقدار بدون برش انتخاب شده و جریمه کامل به خط‌مشی اعمال می‌شود.

۲. تابع هدف با ضریب جریمه تطبیقی (Adaptive KL Penalty): به عنوان رویکرد جایگزین، جریمه واگرایی KL مستقیماً در تابع هدف لحاظ شده و ضریب β به صورت دینامیک به‌روزرسانی می‌شود:

$$L^{\text{KLPEN}}(\theta) = \hat{\mathbb{E}}_t \left[r_t(\theta) \hat{A}_t - \beta\, D_{\text{KL}} \left(\pi_{\theta_{\text{old}}}(\cdot|s_t) \parallel \pi_\theta(\cdot|s_t) \right) \right] \tag{۳۵}$$

در هر تکرار، پس از محاسبه $d = \hat{\mathbb{E}}_t[D_{\text{KL}}(\pi_{\theta_{\text{old}}} \parallel \pi_\theta)]$، ضریب جریمه به این صورت تطبیق می‌یابد:

$$\beta \leftarrow \begin{cases} \beta/2 & \text{if} d < d_{\text{targ}}/1.5 \\ \beta \times 2 & \text{if} d > d_{\text{targ}} \times 1.5 \end{cases} \tag{۳۶}$$

برآوردگر مزیت تعمیم‌یافته (GAE) و تابع هزینه یکپارچه

برای کاهش واریانس، از برآوردگر مزیت تعمیم‌یافته (GAE) منقطع [۳۷] استفاده می‌شود:

$$\hat{A}_t = \sum_{l=0}^{T-t-1} (\gamma\lambda)^l \delta_{t+l}^V \quad \text{where} \quad \delta_t^V = r_t + \gamma V(s_{t+1}) - V(s_t) \tag{۳۷}$$

برای شبکه‌های دارای پارامترهای مشترک میان خط‌مشی و تابع ارزش، تابع هزینه سراسری بهینه‌سازی زیر بیشینه می‌گردد:

$$L_t^{\text{CLIP+VF+S}}(\theta) = \hat{\mathbb{E}}_t \left[L_t^{\text{CLIP}}(\theta) - c_1 \left(V_\theta(s_t) - V_t^{\text{targ}} \right)^2 + c_2 S[\pi_\theta](s_t) \right] \tag{۳۸}$$

که در آن $S[\pi_\theta]$ آنتروپی شانون برای تشویق به اکتشاف (Exploration) و c_1, c_2 ضرایب وزنی هستند.

تحلیل‌های تئوری اطلاعات و آمار

بر اساس بسط تیلور مرتبه دوم واگرایی KL حول نقطه θ_{old} داریم:

$$D_{\text{KL}}(\pi_{\theta_{\text{old}}} \parallel \pi_\theta) \approx \frac{1}{2}(\theta - \theta_{\text{old}})^T F (\theta - \theta_{\text{old}}) \tag{۳۹}$$

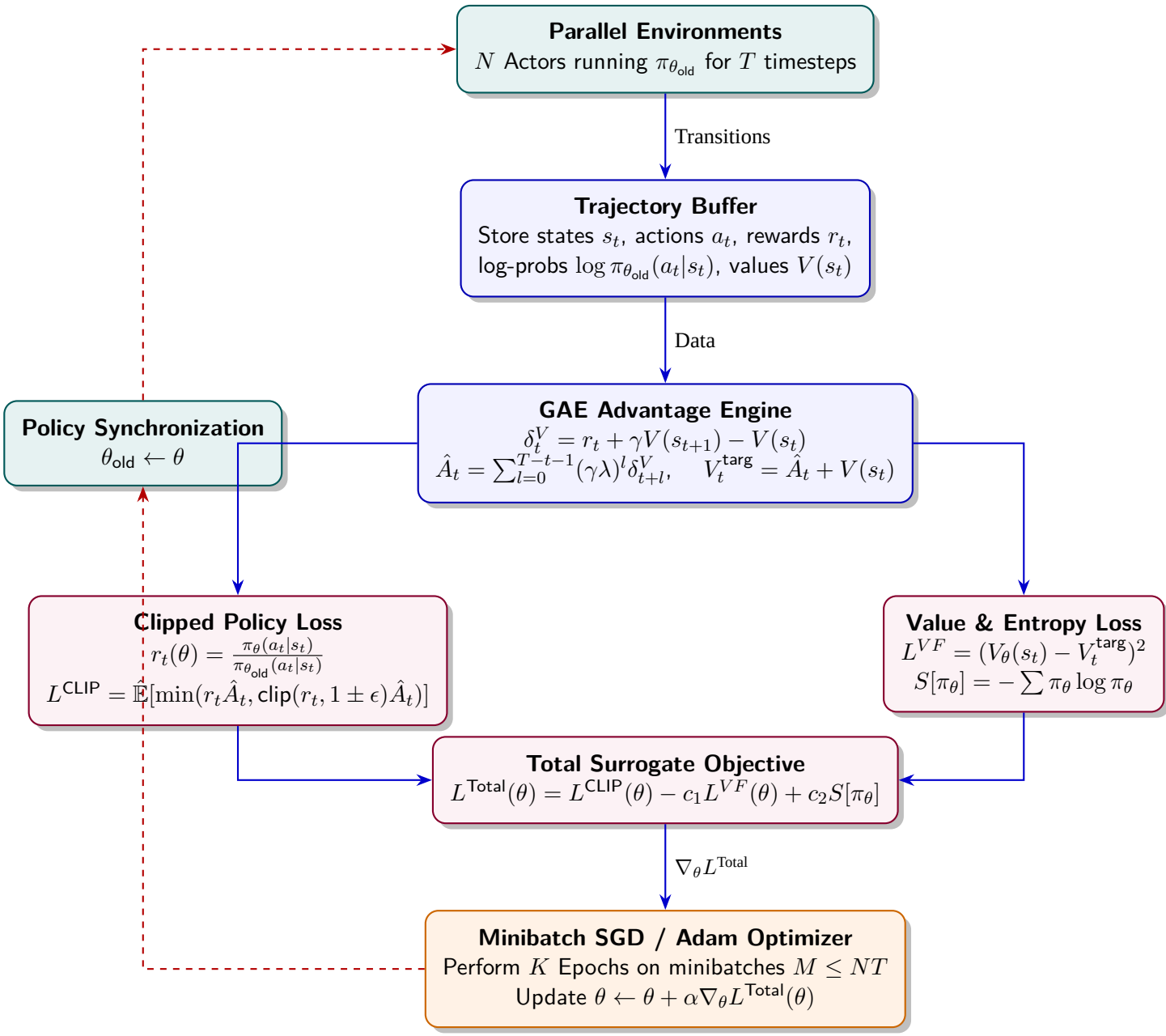
که در آن $F = \mathbb{E}[\nabla_\theta \log \pi_\theta \nabla_\theta \log \pi_\theta^T]$ ماتریس اطلاعات فیشر است. تابع برش در **PPO** تضمین می‌کند که فاصله روی منیفلد ریمانی توزیع‌ها بدون نیاز به تشکیل و معکوس کردن صریح ماتریس F در محدوده‌ای امن محبوس بماند. همچنین با محدود کردن $r_t(\theta) \in [1-\epsilon, 1+\epsilon]$، واریانس بازوزن‌دهی نمونه‌برداری ترجیحی ($\text{Var}_Q[P/Q \cdot f]$) کنترل شده و از انفجار واریانس در بهینه‌سازی چند دوره‌ای جلوگیری می‌شود.

دیاگرام معماری و جریان داده‌های الگوریتم

شکل ۱، ساختار تعامل با محیط، محاسبه مزیت، تشکیل توابع هدف و چرخه به‌روزرسانی چند دوره‌ای در الگوریتم **PPO** را نشان می‌دهد.

شبه‌کد الگوریتم PPO

شبه‌کد استاندارد ساختار Actor-Critic الگوریتم **PPO** در الگوریتم ۲ آورده شده است:



شکل ۱: دیاگرام جامع جریان داده‌ها، محاسبه مزیت و بهینه‌سازی چند دوره‌ای در الگوریتم **PPO**.

Algorithm 2 Proximal Policy Optimization (PPO), Actor-Critic Style

Input: Initial policy parameters θ_0, initial value function parameters ϕ_0, clipping threshold ϵ, GAE parameters γ, λ, epoch count K, minibatch size M, coefficients c_1, c_2.
for iteration $= 1, 2, \ldots$ **do**
  for actor $= 1, \ldots, N$ **do**
    Run policy $\pi_{\theta_{\text{old}}}$ in environment for T timesteps.
    Compute predictions $V_\phi(s_t)$ and log-probabilities $\log \pi_{\theta_{\text{old}}}(a_t|s_t)$.
    Compute TD residuals: $\delta_t^V = r_t + \gamma V_\phi(s_{t+1}) - V_\phi(s_t)$.
    Compute GAE estimates: $\hat{A}_t = \sum_{l=0}^{T-t-1} (\gamma\lambda)^l \delta_{t+l}^V$.
    Compute state-value targets: $V_t^{\text{targ}} = \hat{A}_t + V_\phi(s_t)$.
  end for
  Collect all NT transitions: $\mathcal{D} = \{(s_t, a_t, r_t, \log \pi_{\theta_{\text{old}}}(a_t|s_t), \hat{A}_t, V_t^{\text{targ}})\}$.
  Normalize advantage values: $\hat{A}_t \leftarrow \frac{\hat{A}_t - \mu(\hat{A})}{\sigma(\hat{A}) + 10^{-8}}$.
  for epoch $= 1, \ldots, K$ **do**
    Shuffle dataset $\mathcal{D}$ and partition into minibatches of size M.
    for each minibatch $\mathcal{B} \subset \mathcal{D}$ **do**
      Compute probability ratio: $r_t(\theta) = \exp\left(\log \pi_\theta(a_t|s_t) - \log \pi_{\theta_{\text{old}}}(a_t|s_t)\right)$.
      Compute policy surrogate loss:
        $L^{\text{CLIP}}(\theta) = \frac{1}{|\mathcal{B}|} \sum_{t \in \mathcal{B}} \min\left(r_t(\theta)\hat{A}_t,\ \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\right)$.
      Compute value function squared error loss:
        $L^{\text{VF}}(\phi) = \frac{1}{|\mathcal{B}|} \sum_{t \in \mathcal{B}} \left(V_\phi(s_t) - V_t^{\text{targ}}\right)^2$.
      Compute entropy bonus: $S[\pi_\theta] = \frac{1}{|\mathcal{B}|} \sum_{t \in \mathcal{B}} \mathbb{H}(\pi_\theta(\cdot|s_t))$.
      Maximize total objective using Adam:
        $\theta, \phi \leftarrow \text{Adam}\left(\nabla_{\theta,\phi}\left[L^{\text{CLIP}}(\theta) - c_1 L^{\text{VF}}(\phi) + c_2 S[\pi_\theta]\right]\right)$.
    end for
  end for
  $\theta_{\text{old}} \leftarrow \theta$
end for

نتایج تجربی و ارزیابی بنچمارک‌ها

کارایی **PPO** در مقایسه با انواع روش‌های بهینه‌سازی در جدول ۲ و روی بازی‌های آتاری در جدول ۳ خلاصه شده است:

ابرپارامترهای استاندارد گزارش‌شده در مقاله

جداول ۴ و ۵ مقادیر دقیق ابرپارامترهای بهینه‌شده برای محیط‌های مختلف را نمایش می‌دهند:

جمع‌بندی

الگوریتم **PPO** به واسطه تابع برش بدبینانه L^{CLIP}، بدون نیاز به محاسبات سنگین ماتریس فیشر و با استفاده مستقیم از گرادیان مرتبه اول Adam، رفتار ناحیه اطمینان را شبیه‌سازی کرده و امروزه به عنوان پایه‌ای‌ترین و مطمئن‌ترین الگوریتم یادگیری تقویتی در زمینه‌های مختلف از کنترل حرکت ربات‌های سه‌بعدی [۳۸] تا هم‌ترازسازی مدل‌های زبانی بزرگ از طریق بازخورد انسانی (RLHF) شناخته می‌شود.

جدول ۲: مقایسه عملکرد انواع توابع هدف جایگزین در ۷ محیط پیوسته MuJoCo (برگرفته از مقاله مرجع [۳۰]).

فرمول‌بندی الگوریتم / تابع هدف	میانگین امتیاز نرمال‌شده
بدون برش و بدون جریمه (No clipping or penalty)	-۳۹.۰
برش با $\epsilon = 0.1$ (Clipping)	۷۶.۰
برش با $\epsilon = 0.2$ (Clipping)	**۸۲.۰**
برش با $\epsilon = 0.3$ (Clipping)	۷۰.۰
جریمه تطبیقی با $d_{\text{targ}} = 0.003$ (Adaptive KL)	۶۸.۰
جریمه تطبیقی با $d_{\text{targ}} = 0.01$ (Adaptive KL)	۷۴.۰
جریمه تطبیقی با $d_{\text{targ}} = 0.03$ (Adaptive KL)	۷۱.۰
جریمه ثابت با $\beta = 0.3$ (Fixed KL)	۶۲.۰
جریمه ثابت با $\beta = 1.0$ (Fixed KL)	۷۱.۰
جریمه ثابت با $\beta = 3.0$ (Fixed KL)	۷۲.۰
جریمه ثابت با $\beta = 10.0$ (Fixed KL)	۶۹.۰

جدول ۳: تعداد بازی‌های برده‌شده توسط هر الگوریتم در بنچمارک ۴۹ تایی آتاری (ALE) [۳۰].

معیار ارزیابی	A2C	ACER	PPO	تساوی
(۱) میانگین پاداش در کل طول آموزش	۱	۱۸	**۳۰**	۰
(۲) میانگین پاداش در ۱۰۰ اپیزود پایانی	۱	**۲۸**	۱۹	۱

جدول ۴: ابرپارامترهای **PPO** در بنچمارک رباتیک پیوسته MuJoCo (۱ میلیون گام).

ابرپارامتر	مقدار
افق زمانی داده‌برداری (T)	۲۰۴۸
نرخ یادگیری بهینه‌ساز Adam	3×10^{-4}
تعداد دور بهینه‌سازی در هر آپدیت (K)	۱۰
اندازه دسته کوچک (M)	۶۴
ضریب تخفیف پاداش (γ)	۹۹.۰
پارامتر برآوردگر مزیت (λ)	۹۵.۰

جدول ۵: ابرپارامترهای **PPO** در بنچمارک بازی‌های آتاری (Atari).

ابرپارامتر	مقدار
افق زمانی داده‌برداری (T)	۱۲۸
نرخ یادگیری Adam	$2.5 \times 10^{-4} \times \alpha$
تعداد عامل‌های موازی (N)	۸
تعداد دورها (K)	۳
اندازه دسته کوچک (M)	$32 \times 8 = 256$
ضریب تخفیف پاداش (γ)	۹۹.۰
پارامتر λ در GAE	۹۵.۰
آستانه برش (ϵ)	$0.1 \times \alpha$
ضریب خطای تابع ارزش (c_1)	۰.۱
ضریب پاداش آنتروپی (c_2)	۰۱.۰

۲.۷ الگوریتم بهینه‌سازی نسبی خط‌مشی گروهی (Group Relative Policy Optimization - GRPO)

الگوریتم **GRPO** که توسط تیم DeepSeek-AI در سال ۲۰۲۴ معرفی شد [۳۹]، یکی از پیشرفته‌ترین و کارآمدترین الگوریتم‌های یادگیری تقویتی در حوزه هم‌ترازسازی مدل‌های زبانی بزرگ (LLMs) برای مسائل استدلالی پیچیده و ریاضیات به شمار می‌رود. این الگوریتم با بازتعریف شیوه محاسبه مزیت و حذف کامل مدل منتقد (Critic Model)، بار حافظه و محاسبات الگوریتم‌های سنتی نظیر **PPO** را به طور چشمگیری کاهش داده و همگرایی پایدارتری را به نمایش می‌گذارد.

تاریخچه، انگیزه و ضرورت پیدایش الگوریتم

الگوریتم بهینه‌سازی تقریبی خط‌مشی (**PPO**) [۳۰] به عنوان استاندارد اصلی در مرحله یادگیری تقویتی از بازخورد انسانی (RLHF) [۴۰] شناخته می‌شود. با این حال، استفاده از ساختار کنشگر-منتقد (Actor-Critic) در مدل‌های زبانی با موانع اساسی زیر همراه است:

- **بار سنگین حافظه و محاسبات (Memory and Compute Footprint):** در **PPO** استاندارد، مدل منتقد (V_ψ) از نظر ابعاد پارامتری هم‌اندازه مدل خط‌مشی (π_θ) است. بارگذاری، محاسبه گرادیان و نگهداری متغیرهای بهینه‌ساز برای هر دو مدل، حجم عظیمی از حافظه پردازشی (VRAM) را اشغال می‌کند.
- **پاداش‌های تنک و عدم انطباق با تقریب ارزش توکنی (Sparse Reward Mismatch):** در استدلال‌های ریاضی، پاداش‌ها معمولاً در انتهای توالی (Outcome) یا در انتهای هر گام استدلال (Process) صادر می‌شوند. آموزش یک تابع ارزش برای پیش‌بینی ارزش تک‌تک توکن‌ها در طول یک متن طولانی، دارای واریانس بسیار بالا و خطای تقریب شدید است.
- **طبیعت مقایسه‌ای مدل‌های پاداش (Comparative Nature of Reward Models):** مدل‌های پاداش ذاتاً بر اساس مقایسه نسبی بین گزینه‌ها آموزش می‌بینند. تلاش مدل منتقد برای یادگیری ارزش مطلق به صورت نقطه‌ای، همخوانی کاملی با ماهیت نسبی سیگنال‌های پاداش ندارد.

الگوریتم **GRPO** با حذف کامل مدل ارزش و جایگزینی آن با خط مبنای آماری حاصل از گروهی از پاسخ‌های تولیدشده برای یک پرسش یکسان، این چالش‌ها را به طور بنیادین برطرف می‌سازد.

مبانی نظری و اشتقاق‌های کامل ریاضی

فرض کنید q یک پرسش از توزیع داده‌های آموزشی باشد. الگوریتم **GRPO** به ازای هر پرسش q، گروهی متشکل از G خروجی $\{o_1, o_2, \ldots, o_G\}$ را از خط‌مشی پیشین $\pi_{\theta_{\text{old}}}$ نمونه‌برداری می‌کند.

۱. تابع هدف بهینه‌سازی GRPO: تابع هدف سراسری این الگوریتم به صورت زیر فرمول‌بندی می‌شود:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{\substack{q\sim P(Q) \\ \{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(O|q)}} \left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left(\min\left(\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|q,o_{i,<t})}\hat{A}_{i,t}, \right.\right.\right.$$
$$\left. \text{clip}\left(\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|q,o_{i,<t})}, 1-\varepsilon, 1+\varepsilon \right)\hat{A}_{i,t} \right)$$
$$\left.\left. - \beta\,\mathbb{D}_{\text{KL}}\left(\pi_\theta \parallel \pi_{\text{ref}}\right) \right) \right] \tag{۴۰}$$

که در آن ε آستانه برش برای جلوگیری از به‌روزرسانی‌های مخرب، β ضریب جریمه واگرایی کولبک-لایبلر و $\hat{A}_{i,t}$ مزیت نرمال‌شده در سطح گروه است.

۲. اثبات ناریبی و نامنفی بودن تخمین‌گر واگرایی KL: در این الگوریتم، به جای تزریق جریمه واگرایی به عنوان پاداش به ازای هر توکن، از تخمین‌گر ناریب شولمن [۴۱] به صورت مستقیم در تابع زیان استفاده می‌شود:

$$\mathbb{D}_{\mathsf{KL}}(\pi_\theta \parallel \pi_{\mathsf{ref}}) = \frac{\pi_{\mathsf{ref}}(o_{i,t}|q, o_{i,<t})}{\pi_\theta(o_{i,t}|q, o_{i,<t})} - \log \frac{\pi_{\mathsf{ref}}(o_{i,t}|q, o_{i,<t})}{\pi_\theta(o_{i,t}|q, o_{i,<t})} - 1 \tag{۴۱}$$

اثبات نامنفی بودن: فرض کنید $r = \frac{\pi_{\mathsf{ref}}}{\pi_\theta} > 0$ باشد. تابع $f(r) = r - \log r - 1$ را در نظر می‌گیریم. مشتقات اول و دوم این تابع عبارتند از:

$$f'(r) = 1 - \frac{1}{r}, \qquad f''(r) = \frac{1}{r^2} > 0 \quad (\forall r > 0) \tag{۴۲}$$

از آنجا که مشتق دوم همواره مثبت است، تابع اکیداً محدب بوده و نقطه بحرانی آن در $f'(r) = 0 \implies$ $r = 1$ یک کمینه مطلق است:

$$f(1) = 1 - \log(1) - 1 = 0 \implies f(r) \geq 0 \quad (\forall r > 0) \tag{۴۳}$$

اثبات ناریبی: امید ریاضی این تخمین‌گر تحت توزیع کنونی خط‌مشی π_θ برابر است با:

$$\begin{aligned}
\mathbb{E}_{x\sim\pi_\theta}\left[\frac{\pi_{\mathsf{ref}}(x)}{\pi_\theta(x)} - \log\frac{\pi_{\mathsf{ref}}(x)}{\pi_\theta(x)} - 1\right] &= \sum_x \pi_\theta(x)\left(\frac{\pi_{\mathsf{ref}}(x)}{\pi_\theta(x)} - \log\frac{\pi_{\mathsf{ref}}(x)}{\pi_\theta(x)} - 1\right) \\
&= \sum_x \pi_{\mathsf{ref}}(x) - \sum_x \pi_\theta(x)\log\frac{\pi_{\mathsf{ref}}(x)}{\pi_\theta(x)} - \sum_x \pi_\theta(x) \\
&= 1 + \sum_x \pi_\theta(x)\log\frac{\pi_\theta(x)}{\pi_{\mathsf{ref}}(x)} - 1 = D_{\mathsf{KL}}(\pi_\theta \parallel \pi_{\mathsf{ref}})
\end{aligned} \tag{۴۴}$$

۳. اشتقاق گرادیان و ضریب گرادیان (Gradient Coefficient): با در نظر گرفتن شرایط خط‌مشی در گام نخست ($\pi_{\theta_{\mathsf{old}}} \approx \pi_\theta$)، مشتق تابع هدف را نسبت به پارامترهای θ به دست می‌آوریم:

$$\nabla_\theta \mathcal{L}_{i,t}(\theta) = \nabla_\theta\left[\frac{\pi_\theta(o_{i,t})}{\pi_{\theta_{\mathsf{old}}}(o_{i,t})}\hat{A}_{i,t}\right] - \beta\nabla_\theta\left[\frac{\pi_{\mathsf{ref}}(o_{i,t})}{\pi_\theta(o_{i,t})} - \log\frac{\pi_{\mathsf{ref}}(o_{i,t})}{\pi_\theta(o_{i,t})} - 1\right] \tag{۴۵}$$

با استفاده از تساوی $\nabla_\theta \pi_\theta = \pi_\theta \nabla_\theta \log \pi_\theta$:

$$\nabla_\theta\left[\frac{\pi_\theta}{\pi_{\theta_{\mathsf{old}}}}\hat{A}_{i,t}\right] = \hat{A}_{i,t}\nabla_\theta \log \pi_\theta(o_{i,t}) \tag{۴۶}$$

$$\nabla_\theta\left[\frac{\pi_{\mathsf{ref}}}{\pi_\theta} - \log\pi_{\mathsf{ref}} + \log\pi_\theta - 1\right] = -\frac{\pi_{\mathsf{ref}}}{\pi_\theta^2}\nabla_\theta\pi_\theta + \frac{1}{\pi_\theta}\nabla_\theta\pi_\theta = \left(1 - \frac{\pi_{\mathsf{ref}}}{\pi_\theta}\right)\nabla_\theta \log\pi_\theta(o_{i,t}) \tag{۴۷}$$

با تجمیع دو عبارت فوق، گرادیان کامل به صورت زیر حاصل می‌شود:

$$\nabla_\theta \mathcal{J}_{\mathsf{GRPO}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left(\hat{A}_{i,t} + \beta\left(\frac{\pi_{\mathsf{ref}}(o_{i,t})}{\pi_\theta(o_{i,t})} - 1\right)\right)\nabla_\theta \log\pi_\theta(o_{i,t}|q, o_{i,<t})\right] \tag{۴۸}$$

بنابراین ضریب گرادیان (Gradient Coefficient - GC) در این روش برابر است با:

$$GC_{\mathsf{GRPO}}(q, o, t, \pi_{\theta_{\mathsf{rm}}}) = \hat{A}_{i,t} + \beta\left(\frac{\pi_{\mathsf{ref}}(o_{i,t}|q, o_{i,<t})}{\pi_\theta(o_{i,t}|q, o_{i,<t})} - 1\right) \tag{۴۹}$$

رویکردهای نظارت و محاسبه مزیت در GRPO

الگوریتم **GRPO** امکان بهره‌گیری از دو نوع ناظر را برای تعیین پاداش فراهم می‌سازد:

۱. نظارت بر خروجی (Outcome Supervision RL with GRPO): مدل پاداش به ازای هر پاسخ کامل o_i یک امتیاز r_i تولید می‌کند. بردار پاداش گروهی $\mathbf{r} = \{r_1, r_2, \ldots, r_G\}$ به شکل زیر نرمال‌سازی می‌شود:

$$\hat{A}_{i,t} = \tilde{r}_i = \frac{r_i - \text{mean}(\mathbf{r})}{\text{std}(\mathbf{r}) + 10^{-8}}, \quad \forall t \in \{1, \ldots, |o_i|\} \tag{۵۰}$$

۲. نظارت بر فرآیند (Process Supervision RL with GRPO): مدل پاداش فرایندی (PRM) [۴۲، ۴۳] به انتهای هر گام استدلال j پاداش $r_i^{\text{index}(j)}$ اختصاص می‌دهد. با نرمال‌سازی کل پاداش‌های مرحله‌ای درون گروه:

$$\tilde{r}_i^{\text{index}(j)} = \frac{r_i^{\text{index}(j)} - \text{mean}(\mathbf{R})}{\text{std}(\mathbf{R}) + 10^{-8}} \tag{۵۱}$$

سپس مزیت هر توکن در موقعیت t از مجموع پاداش‌های گام‌های پسین محاسبه می‌شود:

$$\hat{A}_{i,t} = \sum_{\substack{j=1 \\ \text{index}(j) \geq t}}^{K_i} \tilde{r}_i^{\text{index}(j)} \tag{۵۲}$$

چارچوب یکپارچه روش‌های هم‌ترازسازی

مقاله یک صورت‌بندی فراگیر ارائه می‌دهد که تمامی روش‌های تنظیم دقیق و یادگیری تقویتی را تحت معادله گرادیان زیر متحد می‌سازد:

$$\nabla_\theta \mathcal{J}_{\mathcal{A}}(\theta) = \mathbb{E}_{(q,o)\sim\mathcal{D}} \left[\frac{1}{|o|} \sum_{t=1}^{|o|} GC_{\mathcal{A}}(q, o, t, \pi_{\text{rf}}) \nabla_\theta \log \pi_\theta(o_t | q, o_{<t}) \right] \tag{۵۳}$$

جدول ۶ مقایسه ساختاری این روش‌ها را نشان می‌دهد:

جدول ۶: تحلیل ساختار روش‌های هم‌ترازسازی در چارچوب یکپارچه گرادیان [۳۹].

روش (Method)	منبع داده ($\mathcal{D}$)	تابع پاداش (π_{rf})	ضریب گرادیان (GC)	نوع نمونه‌برداری
SFT	$q, o \sim P_{\text{sft}}(Q, O)$	انتخاب انسانی	1	آفلاین ایستا
RFT [۴۴]	$q \sim P_{\text{sft}}(Q),\ o \sim \pi_{\text{sft}}(O\|q)$	قاعده قطعی (Rule)	$\mathbb{I}(o \text{correct is })$	آفلاین تصادفی
DPO [۴۵]	$q \sim P_{\text{sft}}(Q),\ o^+, o^- \sim \pi_{\text{sft}}(O\|q)$	ترجیح / قاعده	$\sigma\left(\beta \log \frac{\pi_\theta(o^-)}{\pi_{\text{ref}}(o^-)} - \beta \log \frac{\pi_\theta(o^+)}{\pi_{\text{ref}}(o^+)}\right)$	آفلاین زوجی
Online RFT	$q \sim P_{\text{sft}}(Q),\ o \sim \pi_\theta(O\|q)$	قاعده قطعی (Rule)	$\mathbb{I}(o \text{correct is })$	آنلاین تصادفی
PPO [۳۰]	$q \sim P_{\text{sft}}(Q),\ o \sim \pi_\theta(O\|q)$	مدل پاداش	$A_t = \text{GAE}(\{r_{\geq t}\}, V_\psi)$	آنلاین پیوسته
GRPO	$q \sim P_{\text{sft}}(Q),\ \{o_i\}_{i=1}^G \sim \pi_\theta(O\|q)$	مدل پاداش / فرآیند	$\hat{A}_{i,t} + \beta\left(\frac{\pi_{\text{ref}}(o_{i,t})}{\pi_\theta(o_{i,t})} - 1\right)$	**آنلاین گروهی**

تحلیل‌های تئوری اطلاعات، آمار و چرایی موفقیت الگوریتم

- **کاهش واریانس بدون منتقد:** میانگین پاداش درون‌گروهی $b(q) = \frac{1}{G}\sum_{i=1}^G r_i$ به عنوان خط مبنای نااریب عمل کرده و واریانس تخمین‌گر به میزان $\mathcal{O}(1/G)$ کاهش می‌یابد بدون آنکه نیازی به خطای تخمین مدل منتقد ($\epsilon_{\text{critic}} = V^*(s) - V_\psi(s)$) باشد.

- **تحلیل معیارهای Pass@K در برابر Maj@K:** ارزیابی‌های تجربی روی بنچمارک‌های GSM8K و MATH نشان می‌دهند که یادگیری تقویتی توان بالقوه خام (Pass@K) مدل را نسبت به مرحله SFT تغییر چشمگیری نمی‌دهد؛ بلکه توزیع احتمالاتی پاسخ‌ها را کالیبره کرده و با متمرکز ساختن چگالی احتمال حول پاسخ‌های صحیح، دقت رأی‌گیری اکثریت (Maj@K) و پاسخ برتر (Top-1) را به شدت ارتقا می‌بخشد.

دیاگرام معماری و جریان داده‌های الگوریتم

شکل ۲ ساختار داده‌برداری گروهی، نرمال‌سازی آماری، محاسبه ضریب گرادیان و چرخه به‌روزرسانی بدون منتقد را در الگوریتم **GRPO** نمایش می‌دهد.

شبه‌کد الگوریتم GRPO

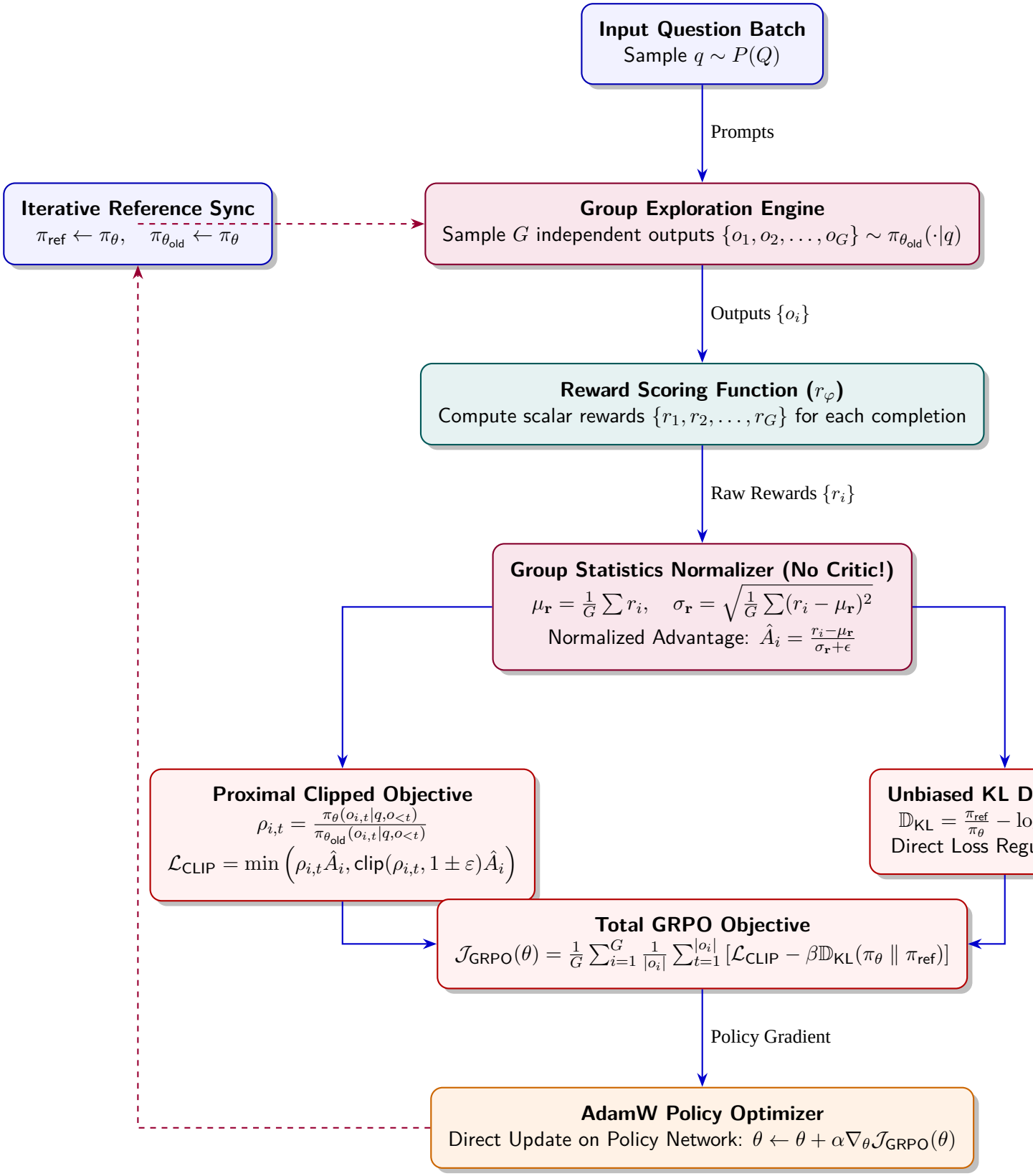
شبه‌کد کامل فرایند بهینه‌سازی سیاست نسبی گروهی تکرارشونده در الگوریتم ۳ تشریح شده است:

Algorithm 3 Iterative Group Relative Policy Optimization (GRPO)

Input: Initial policy $\pi_{\theta_{\text{init}}}$, reward model r_φ, prompt dataset $\mathcal{D}$, hyperparameters $\varepsilon, \beta, \mu, G$, iterations I, steps M.
Initialize policy $\pi_\theta \leftarrow \pi_{\theta_{\text{init}}}$
for iteration $= 1, \ldots, I$ **do**
  Set reference policy $\pi_{\text{ref}} \leftarrow \pi_\theta$
  for step $= 1, \ldots, M$ **do**
    Sample batch of prompts $\mathcal{D}_b \sim \mathcal{D}$
    Update old policy $\pi_{\theta_{\text{old}}} \leftarrow \pi_\theta$
    Sample G candidate outputs $\{o_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot|q)$ for each prompt $q \in \mathcal{D}_b$
    Compute rewards $\{r_i\}_{i=1}^G$ using reward model r_φ
    Compute group mean and standard deviation:
      $\mu_{\mathbf{r}} = \frac{1}{G}\sum_{i=1}^G r_i, \quad \sigma_{\mathbf{r}} = \sqrt{\frac{1}{G}\sum_{i=1}^G (r_i - \mu_{\mathbf{r}})^2 + 10^{-8}}$
    Compute group relative advantages: $\hat{A}_{i,t} = \frac{r_i - \mu_{\mathbf{r}}}{\sigma_{\mathbf{r}}}$
    for GRPO iteration $= 1, \ldots, \mu$ **do**
      Compute ratio: $\rho_{i,t} = \frac{\pi_\theta(o_{i,t}|q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|q, o_{i,<t})}$
      Compute clipped advantage objective:
        $\mathcal{L}_{i,t}^{\text{CLIP}} = \min\left(\rho_{i,t}\hat{A}_{i,t},\ \text{clip}(\rho_{i,t}, 1-\varepsilon, 1+\varepsilon)\hat{A}_{i,t}\right)$
      Compute token-level Schulman KL penalty:
        $\mathbb{D}_{\text{KL},i,t} = \frac{\pi_{\text{ref}}(o_{i,t}|q, o_{i,<t})}{\pi_\theta(o_{i,t}|q, o_{i,<t})} - \log\frac{\pi_{\text{ref}}(o_{i,t}|q, o_{i,<t})}{\pi_\theta(o_{i,t}|q, o_{i,<t})} - 1$
      Maximize GRPO loss with AdamW:
        $\theta \leftarrow \theta + \alpha\nabla_\theta\left[\frac{1}{G|\mathcal{D}_b|}\sum_{q\in\mathcal{D}_b}\sum_{i=1}^G \frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left(\mathcal{L}_{i,t}^{\text{CLIP}} - \beta\mathbb{D}_{\text{KL},i,t}\right)\right]$
    end for
    Update reward model r_φ through replay mechanism with 10% historical samples.
  end for
end for
Output: Optimized Policy π_θ

نتایج تجربی و ارزیابی بنچمارک‌ها

نتایج حاصل از اعمال الگوریتم **GRPO** بر روی مدل پایه **DeepSeekMath-Instruct 7B** در مقایسه با سایر مدل‌های باز و بسته در جدول ۷ ارائه شده است:



شکل ۲: دیاگرام جامع جریان داده‌ها، ارزیابی آماری گروهی و به‌روزرسانی بدون منتقد در الگوریتم **GRPO**.

جدول ۷: مقایسه عملکرد **DeepSeekMath-RL (GRPO)** با مدل‌های پیشرو در بنچمارک‌های استدلال ریاضی [۳۹].

مدل (Model)	تعداد پارامتر	GSM8K (CoT)	MATH (CoT)	MGSM-zh	CMATH
مدل‌های تجاری و بسته (Closed-Source Models)					
Gemini Ultra	-	۴%.۹۴	۲%.۵۳	-	-
GPT-4	-	۰%.۹۲	۹%.۵۲	-	۰%.۸۶
Inflection-2	-	۴%.۸۱	۸%.۳۴	-	-
GLM-4	-	۶%.۸۷	۹%.۴۷	-	-
مدل‌های منبع‌باز (Open-Source Models)					
InternLM2-Math	۲۰B	۶%.۸۲	۷%.۳۷	-	-
Qwen	۷۲B	۹%.۷۸	۲%.۳۵	-	-
WizardMath-v1.1	۷B	۲%.۸۳	۰%.۳۳	-	-
DeepSeekMath-Instruct	۷B	۹%.۸۲	۸%.۴۶	۲%.۷۳	۶%.۸۴
DeepSeekMath-RL (GRPO)	**۷B**	**۲%.۸۸**	**۷%.۵۱**	**۶%.۷۹**	**۸%.۸۸**

ابرپارامترهای استاندارد گزارش‌شده در مقاله

جدول ۸ پیکربندی ابرپارامترهای بهینه پیاده‌سازی **GRPO** را نشان می‌دهد:

جدول ۸: ابرپارامترهای آموزش یادگیری تقویتی با **GRPO** در مدل **DeepSeekMath 7B**.

ابرپارامتر (Hyperparameter)	مقدار
نرخ یادگیری مدل خط‌مشی (α)	1×10^{-6}
نرخ یادگیری مدل پاداش اولیه	2×10^{-5}
تعداد نمونه‌های خروجی در هر گروه (G)	۶۴
ضریب جریمه واگرایی کولبک-لایبلر (β)	۰۴.۰
آستانه برش پروگزیمال (ε)	۲.۰
اندازه دسته بهینه‌سازی (Batch Size)	۱۰۲۴
حداکثر طول توالی خروجی (Max Length)	۱۰۲۴
دمای نمونه‌برداری اکتشافی (Sampling Temperature)	۷.۰

جمع‌بندی

الگوریتم **GRPO** با حذف کامل سربار محاسباتی و حافظه مدل منتقد و جایگزینی آن با نرمال‌سازی آماری در سطح گروه‌های هم‌ارز، مسیر نوینی را در یادگیری تقویتی مدل‌های زبانی گشوده است. این نوآوری به مدل **DeepSeekMath 7B** اجازه داد تا با صرفه‌جویی شدید در منابع سخت‌افزاری، به دقت فراتر از ۵۱٪ در بنچمارک معتبر MATH دست یافته و استانداردهای جدیدی در زمینه استدلال هوش مصنوعی پایه‌گذاری کند.

۳.۷ الگوریتم بهینه‌سازی توالی گروهی خط‌مشی (Group Sequence Policy Optimization - GSPO)

الگوریتم **GSPO** که در سال ۲۰۲۵ توسط تیم کوئن (Qwen Team) در شرکت علی‌بابا (Alibaba Inc.) معرفی شد [۴۶]، یک چارچوب نوین، فوق‌العاده پایدار و کارآمد در زمینه یادگیری تقویتی (Reinforcement Learning) برای مدل‌های زبانی بزرگ (LLMs) به شمار می‌رود. این الگوریتم با بازنگری در

شیوه اعمال «نمونه‌برداری اهمیت» (Importance Sampling) و اصلاح خطاهای ساختاری الگوریتم‌های پیشین نظیر **GRPO** [۳۹]، کلیه فرآیندهای وزن‌دهی نسبت درستی‌نمایی، برش گرادیان (Clipping) و بهینه‌سازی را از سطح «توکن» (Token-level) به سطح «توالی کامل» (Sequence-level) منتقل کرده است. این رویکرد به عنوان ستون فقرات یادگیری تقویتی در آموزش مدل‌های استدلالی پیشرفته خانواده **Qwen3** [۴۷، ۴۸] به کار گرفته شده است.

تاریخچه، انگیزه و ضرورت پیدایش الگوریتم

با گسترش مدل‌های زبانی استدلالی نظیر OpenAI o1 [۴۹] و **DeepSeek-R1** [۵۰]، نقش یادگیری تقویتی در مقیاس‌پذیری زمان تفکر (Test-Time Compute) و تولید زنجیره‌های طولانی استدلال (Long Chain-of-Thought) برای حل مسائل ریاضی و برنامه‌نویسی رقابتی اثبات گردید.

پیش از ابداع **GSPO**، روند توسعه الگوریتم‌های یادگیری تقویتی در مدل‌های زبانی با چالش‌های اساسی روبرو بود:

- **محدودیت‌های PPO [۳۰]:** الگوریتم **PPO** نیازمند یک مدل منتقد یا ارزش (Value/Critic Model) مجزا برای برآورد مزیت در هر توکن است. این مدل علاوه بر تحمیل بار حافظه و محاسباتی هم‌اندازه با مدل خط‌مشی اصلی، در پاسخ‌های طولانی استدلالی (چندین هزار توکن) به دلیل انباشت خطای پیش‌بینی، به شدت ناپایدار و غیرقابل اعتماد می‌گردد.
- **ظهور و معضلات الگوریتم GRPO [۳۹]:** الگوریتم **GRPO** با حذف مدل منتقد و جایگزینی آن با برآورد مزیت نسبی در یک گروه از پاسخ‌ها (Group Relative Advantage)، نیاز به حافظه اضافی را رفع کرد. با این حال، **GRPO** در آموزش مدل‌های بسیار بزرگ و در استدلال‌های طولانی، مکرراً با پدیده «فروپاشی فاجعه‌بار و غیرقابل بازگشت مدل» (Irreversible Model Collapse) مواجه می‌شود [۴۷، ۵۱]. در صورت وقوع این فروپاشی، هیچ‌گونه دستکاری در ابرپارامترها یا بارگذاری مجدد چک‌پوینت‌ها کمکی به همگرایی مجدد نخواهد کرد.

نویسندگان مقاله **GSPO** ریشه این ناپایداری در **GRPO** را در دو نقص تئوریک شناسایی کردند:

۱. **نقض مبانی ریاضی نمونه‌برداری اهمیت (Importance Sampling):** اصل نمونه‌برداری اهمیت برای تخمین امید ریاضی یک تابع $f(z)$ تحت توزیع هدف π_{tar} با استفاده از توزیع رفتاری π_{beh} به صورت زیر بیان می‌شود:

$$\mathbb{E}_{z\sim\pi_{\text{tar}}}[f(z)] = \mathbb{E}_{z\sim\pi_{\text{beh}}}\left[\frac{\pi_{\text{tar}}(z)}{\pi_{\text{beh}}(z)}f(z)\right] \approx \frac{1}{N}\sum_{k=1}^{N}\frac{\pi_{\text{tar}}(z^{(k)})}{\pi_{\text{beh}}(z^{(k)})}f(z^{(k)}) \qquad (۵۴)$$

این رابطه ریاضی از نظر آماری تنها در صورتی معتبر و بدون انحراف است که میانگین‌گیری روی تعداد زیادی نمونه ($N \gg 1$) از توزیع رفتاری صورت پذیرد. در حالی که **GRPO** نسبت اهمیت $\frac{\pi_\theta(y_{i,t}|x,y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t}|x,y_{i,<t})}$ را برای هر تک‌توکن t و با تنها یک نمونه تکی ($N = 1$) اعمال می‌کند. این کار به جای تصحیح انحراف توزیع، نویز ضرب‌شونده با واریانس بسیار بالا را وارد گرادیان کرده که با طولانی‌تر شدن زنجیره استدلال انباشته شده و مدل را نابود می‌سازد.

۲. **عدم انطباق واحد بهینه‌سازی و واحد پاداش (Mismatch Principle):** پاداش ارزیاب $r(x, y)$ به کل توالی پاسخ اعطا می‌شود نه به تک‌تک توکن‌ها. بنابراین، اعمال تصحیح خارج از خط‌مشی (Off-Policy Correction) و عملگر برش در سطح توکن‌های مجزا، فرمول‌بندی نادرست و ناسازگاری ایجاد می‌کند.

مبانی نظری و اشتقاق‌های کامل ریاضی GSPO

در یک مدل زبانی خودبازگشتی، درستی‌نمایی (Likelihood) یک پاسخ $y = (y_1, y_2, \ldots, y_{|y|})$ به ازای پرسش ورودی x تحت خط‌مشی π_θ برابر است با:

$$\pi_\theta(y|x) = \prod_{t=1}^{|y|} \pi_\theta(y_t|x, y_{<t}) \tag{۵۵}$$

الگوریتم **GSPO**، تابع هدف بهینه‌سازی را مستقیماً در سطح توالی فرمول‌بندی می‌کند:

$$\mathcal{J}_{\text{GSPO}}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\{y_i\}_{i=1}^G\sim\pi_{\theta_{\text{old}}}(\cdot|x)} \left[\frac{1}{G}\sum_{i=1}^{G} \min\left(s_i(\theta)\hat{A}_i,\ \text{clip}(s_i(\theta), 1-\varepsilon, 1+\varepsilon)\hat{A}_i\right)\right] \tag{۵۶}$$

که در آن G اندازه گروه پاسخ‌های هم‌زمان به ازای هر پرسش بوده و مزیت گروهی $\hat{A}_i$ بر اساس امتیاز پاداش تاییدکننده $r(x, y_i) \in [0, 1]$ به صورت زیر نرمال‌سازی می‌شود:

$$\hat{A}_i = \frac{r(x, y_i) - \text{mean}\left(\{r(x, y_i)\}_{i=1}^G\right)}{\text{std}\left(\{r(x, y_i)\}_{i=1}^G\right)} \tag{۵۷}$$

نسبت اهمیت توالی $s_i(\theta)$ بر اساس مفهوم درستی‌نمایی توالی با نرمال‌سازی طولی (Length-Normalized Sequence Likelihood) [۵۲] تعریف می‌شود:

$$s_i(\theta) \triangleq \left(\frac{\pi_\theta(y_i|x)}{\pi_{\theta_{\text{old}}}(y_i|x)}\right)^{\frac{1}{|y_i|}} = \exp\left(\frac{1}{|y_i|}\sum_{t=1}^{|y_i|} \log \frac{\pi_\theta(y_{i,t}|x, y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t}|x, y_{i,<t})}\right) \tag{۵۸}$$

اثبات گام‌به‌گام اشتقاق گرادیان GSPO: با صرف‌نظر از عملگر غیرخطی برش جهت بررسی گرادیان بدون مرز، گرادیان تابع هدف نسبت به پارامترهای θ به صورت زیر محاسبه می‌شود:

$$\begin{aligned}\nabla_\theta \mathcal{J}_{\text{GSPO}}(\theta) &= \nabla_\theta\, \mathbb{E}_{x,\{y_i\}_{i=1}^G}\left[\frac{1}{G}\sum_{i=1}^{G} s_i(\theta)\hat{A}_i\right] \\ &= \mathbb{E}_{x,\{y_i\}_{i=1}^G}\left[\frac{1}{G}\sum_{i=1}^{G} \hat{A}_i\, \nabla_\theta s_i(\theta)\right]\end{aligned} \tag{۵۹}$$

با توجه به اینکه برای هر تابع مثبت $f(\theta)$ داریم $\nabla_\theta f(\theta) = f(\theta)\nabla_\theta \log f(\theta)$، مشتق نسبت اهمیت توالی برابر است با:

$$\nabla_\theta s_i(\theta) = s_i(\theta)\, \nabla_\theta \log s_i(\theta) \tag{۶۰}$$

با اعمال لگاریتم بر طرفین معادله (۵۸):

$$\log s_i(\theta) = \frac{1}{|y_i|}\sum_{t=1}^{|y_i|} \left[\log \pi_\theta(y_{i,t}|x, y_{i,<t}) - \log \pi_{\theta_{\text{old}}}(y_{i,t}|x, y_{i,<t})\right] \tag{۶۱}$$

از آنجا که جمله مربوط به θ_{old} مستقل از θ است، مشتق آن صفر می‌شود:

$$\nabla_\theta \log s_i(\theta) = \frac{1}{|y_i|}\sum_{t=1}^{|y_i|} \nabla_\theta \log \pi_\theta(y_{i,t}|x, y_{i,<t}) \tag{۶۲}$$

بنابراین، فرمول بسته نهایی گرادیان **GSPO** حاصل می‌شود:

$$\nabla_\theta \mathcal{J}_{\text{GSPO}}(\theta) = \mathbb{E}_{x,\{y_i\}_{i=1}^G}\left[\frac{1}{G}\sum_{i=1}^{G}\left(\frac{\pi_\theta(y_i|x)}{\pi_{\theta_{\text{old}}}(y_i|x)}\right)^{\frac{1}{|y_i|}}\hat{A}_i \cdot \frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\nabla_\theta \log \pi_\theta(y_{i,t}|x, y_{i,<t})\right] \tag{۶۳}$$

در مقام مقایسه، گرادیان الگوریتم **GRPO** به صورت زیر است:

$$\nabla_\theta \mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{x,\{y_i\}_{i=1}^G}\left[\frac{1}{G}\sum_{i=1}^{G}\hat{A}_i \cdot \frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\left(\frac{\pi_\theta(y_{i,t}|x, y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t}|x, y_{i,<t})}\right)\nabla_\theta \log \pi_\theta(y_{i,t}|x, y_{i,<t})\right] \tag{۶۴}$$

تحلیل مقایسه‌ای دو گرادیان: در **GRPO** (معادله (۶۴))، گرادیان هر توکن با یک وزن نابرابر و تصادفی $\frac{\pi_\theta(y_{i,t})}{\pi_{\theta_{\text{old}}}(y_{i,t})}$ مقیاس‌بندی می‌شود که این نوسانات نامتعادل در طول هزاران گام به فروپاشی مدل می‌انجامد. اما در **GSPO** (معادله (۶۳))، تمامی توکن‌های درون یک پاسخ با یک ضریب کاملاً همگن و یکنواخت ($s_i(\theta)\frac{\hat{A}_i}{|y_i|}$) وزن‌دهی می‌شوند که عامل اصلی ثبات بی‌نظیر این الگوریتم است.

گونه تعمیم‌یافته برای توکن‌ها (GSPO-token Variant)

برای سناریوهایی نظیر یادگیری تقویتی چندنوبته (Multi-turn RL) یا استفاده از مدل‌های پاداش فرآیندی (PRM) که نیازمند مزیت‌های متفاوت برای هر گام یا توکن ($\hat{A}_{i,t}$) هستیم، گونه **GSPO-token** با بهره‌گیری از عملگر سد گرادیان ($\text{sg}[\cdot]$ یا Stop-Gradient) معرفی شده است:

$$\mathcal{J}_{\text{GSPO-token}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\min\left(s_{i,t}(\theta)\hat{A}_{i,t},\ \text{clip}(s_{i,t}(\theta), 1-\varepsilon, 1+\varepsilon)\hat{A}_{i,t}\right)\right] \tag{۶۵}$$

که در آن نسبت توکن با ساختار زیر تعریف می‌شود:

$$s_{i,t}(\theta) \triangleq \text{sg}[s_i(\theta)] \cdot \frac{\pi_\theta(y_{i,t}|x, y_{i,<t})}{\text{sg}\left[\pi_\theta(y_{i,t}|x, y_{i,<t})\right]} \tag{۶۶}$$

از آنجا که عبارت کسری از نظر مقداری دقیقاً برابر با 1 است، مقدار عددی $s_{i,t}(\theta) = s_i(\theta)$ بوده و شرط برش در سطح کل توالی حفظ می‌شود. در زمان مشتق‌گیری، گرادیان مستقیماً روی توکن t اثر می‌گذارد:

$$\nabla_\theta \mathcal{J}_{\text{GSPO-token}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\left(\frac{\pi_\theta(y_i|x)}{\pi_{\theta_{\text{old}}}(y_i|x)}\right)^{\frac{1}{|y_i|}} \cdot \frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\hat{A}_{i,t}\nabla_\theta \log \pi_\theta(y_{i,t}|x, y_{i,<t})\right] \tag{۶۷}$$

در حالتی که مزیت تمام توکن‌ها یکسان باشد ($\hat{A}_{i,t} = \hat{A}_i$)، این گرادیان دقیقاً و به صورت تحلیلی با گرادیان **GSPO** اصلی هم‌ارز می‌گردد.

تحلیل‌های تئوری اطلاعات، آمار و دینامیک مدل‌های MoE

۱. کنترل واریانس و واگرایی KL: بر اساس تئوری اطلاعات، نسبت درستی‌نمایی ضرب‌شونده بدون نرمال‌سازی طولی دارای واریانسی متناسب با طول پاسخ است ($\text{Var} \propto \mathcal{O}(|y|)$). با اعمال میانگین هندسی در توان $\frac{1}{|y|}$، واریانس نسبت اهمیت به صورت $\mathcal{O}(1/|y|)$ میرا شده و کران واگرایی اطلاعاتی میان سیاست جدید و قدیم کنترل می‌شود.

۲. **تثبیت دینامیک آموزش در معماری‌های ترکیبی از متخصصان (MoE):** در مدل‌های MoE نظیر **Qwen3-30B-A3B** با ۴۸ لایه، پس از هر به‌روزرسانی گرادیان، تخصیص حدود ٪۱۰ از روترهای متخصصان تغییر می‌کند (Expert-Activation Volatility). این پدیده باعث نوسان شدید در درستی‌نمایی محلی توکن‌ها شده و آموزش **GRPO** را به شکست قطعی می‌کشاند، مگر آنکه از روش پرهزینه «بازپخش مسیریابی» (Routing Replay) برای اجبار به استفاده از مسیرهای قدیمی استفاده شود. **GSPO** به دلیل تمرکز بر درستی‌نمایی کل توالی و عدم حساسیت به تغییرات جابجایی روترهای محلی، نیاز به Routing Replay را کاملاً منسوخ کرده و ظرفیت کامل مدل را بدون افت سرعت آزاد می‌سازد.

۳. **بهینه‌سازی زیرساخت آموزش (Infrastructure):** تفاوت‌های دقت ممیز شناور میان موتورهای استنتاج (vLLM / SGLang) و موتورهای آموزش (Megatron-LM) در سطح تک‌توکن‌ها خطاساز است. **GSPO** به دلیل پایداری در سطح توالی، به مدل اجازه می‌دهد تا مستقیماً لاگ‌احتمالات خروجی از موتور استنتاج را بدون نیاز به محاسبه مجدد (Recomputation) در موتور آموزش استفاده کند.

دیاگرام معماری و جریان داده‌های الگوریتم

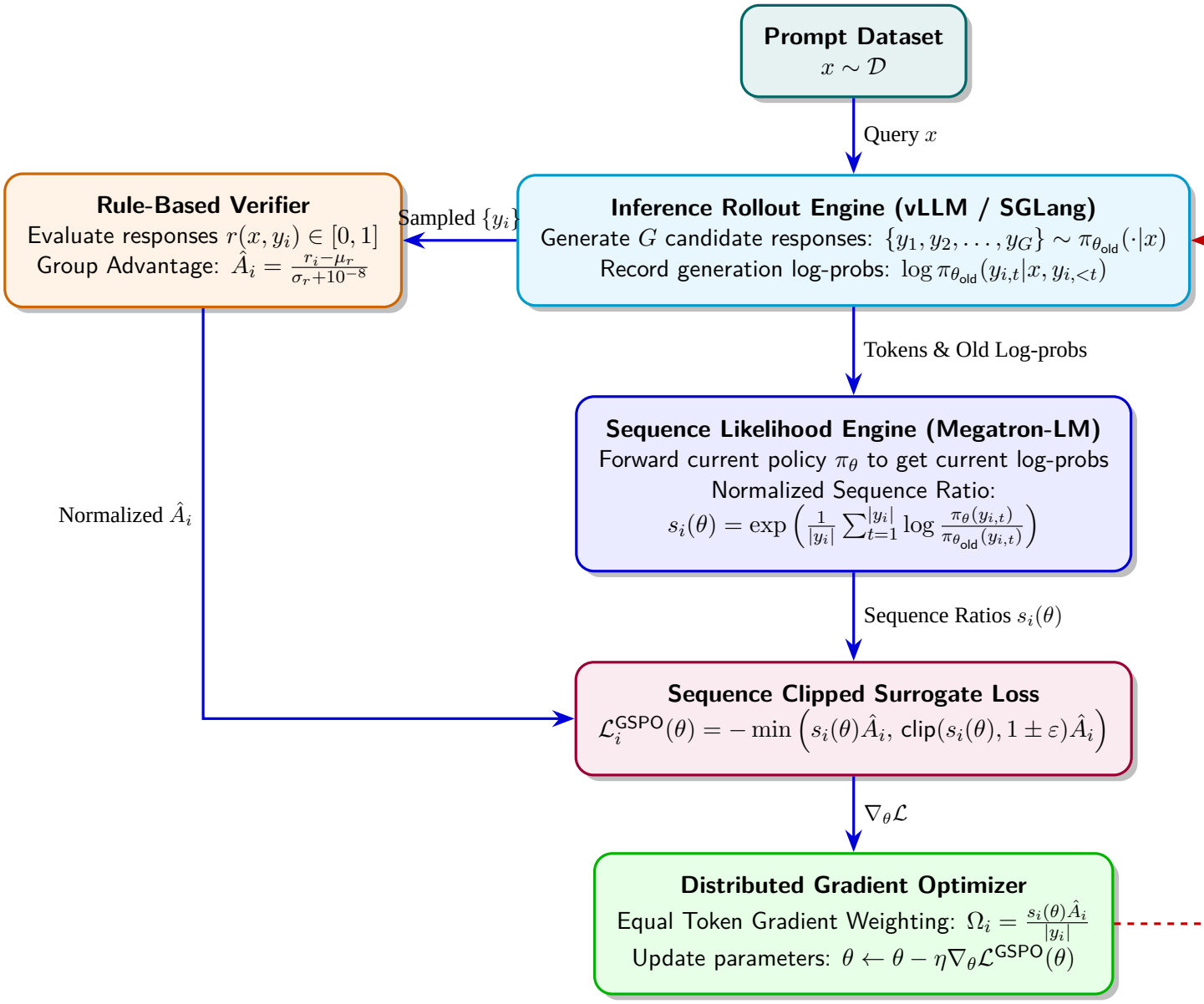
شکل ۳ ساختار جامع خط لوله آموزش، محاسبات نسبت توالی و تفاوت بنیادین جریان داده در الگوریتم **GSPO** در مقایسه با روش‌های پیشین را نمایش می‌دهد.

شبه‌کد الگوریتم GSPO

شبه‌کد رسمی و استاندارد پیاده‌سازی مقیاس‌پذیر الگوریتم **GSPO** در الگوریتم ۴ ارائه شده است:

Algorithm 4 Group Sequence Policy Optimization (GSPO)

Input: Initial actor policy parameters θ_0, verifier reward function $r(x, y)$, group size G, clipping threshold ε, learning rate η, number of mini-batches K.
for iteration $= 1, 2, \ldots$ **do**
  Sample a batch of queries $\{x_j\}_{j=1}^{B} \sim \mathcal{D}$.
  Set behavior policy $\theta_{\text{old}} \leftarrow \theta$.
  Initialize rollout buffer $\mathcal{B} \leftarrow \emptyset$.
  for each query $x \in \{x_j\}_{j=1}^{B}$ **do**
    Generate G responses $\{y_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot|x)$ using inference engine.
    Compute scalar rewards $r_i = r(x, y_i)$ for $i \in \{1, \ldots, G\}$.
    Compute group advantage: $\hat{A}_i = \frac{r_i - \text{mean}(\{r_k\})}{\text{std}(\{r_k\}) + 10^{-8}}$.
    for $i = 1, \ldots, G$ **do**
      Store rollout record $(x, y_i, \hat{A}_i, \{\log \pi_{\theta_{\text{old}}}(y_{i,t}|x, y_{i,<t})\}_{t=1}^{|y_i|})$ in $\mathcal{B}$.
    end for
  end for
  Partition buffer $\mathcal{B}$ into K mini-batches $\{\mathcal{B}_m\}_{m=1}^{K}$.
  for each mini-batch $\mathcal{B}_m$ **do**
    Forward current policy π_θ to get current log-probabilities $\log \pi_\theta(y_{i,t}|x, y_{i,<t})$.
    Compute length-normalized sequence importance ratio for each sample:
      $s_i(\theta) = \exp\left(\frac{1}{|y_i|} \sum_{t=1}^{|y_i|} \left(\log \pi_\theta(y_{i,t}|x, y_{i,<t}) - \log \pi_{\theta_{\text{old}}}(y_{i,t}|x, y_{i,<t})\right)\right)$.
    Compute sequence clipped surrogate loss:
      $\mathcal{L}(\theta) = -\frac{1}{|\mathcal{B}_m|} \sum_{i \in \mathcal{B}_m} \min\left(s_i(\theta)\hat{A}_i,\ \text{clip}(s_i(\theta), 1-\varepsilon, 1+\varepsilon)\hat{A}_i\right)$.
    Perform backward pass and parameter update:
      $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}(\theta)$.
  end for
end for



شکل ۳: دیاگرام جامع معماری خط لوله آموزش، محاسبه نسبت اهمیت توالی با نرمال‌سازی طولی و به‌روزرسانی گرادیان همگن در الگوریتم **GSPO**.

نتایج تجربی و ارزیابی بنچمارک‌ها

ارزیابی تجربی **GSPO** بر روی مدل پایه **Qwen3-30B-A3B-Base** در مقایسه با **GRPO** در بنچمارک‌های استاندارد استدلالی، برنامه‌نویسی و ریاضی در جدول ۹ آورده شده است.

جدول ۹: مقایسه عملکرد و پایداری الگوریتم‌های **GRPO** و **GSPO** بر روی مدل **Qwen3-30B-A3B** (برگرفته از [۴۶]).

الگوریتم یادگیری تقویتی	(Pass@۱) ۲۴' AIME	(Pass@۱) LiveCodeBench	(Elo) CodeForces	پایداری در آموزش MoE
GRPO (بدون Routing Replay)	ناموفق (فروپاشی)	ناموفق (فروپاشی)	ناموفق (فروپاشی)	کاملاً ناپایدار
GRPO (با Routing Replay)	۲۴.۷۷٪	۴۰.۶۳٪	۱۹۲۰	نیازمند سربار حافظه
GSPO (رویکرد پیشنهادی)	**۶۰.۸۲٪**	**۱۱.۶۸٪**	**۲۰۶۵**	**پایداری کامل و ذاتی**

پارادوکس کسر توکن‌های برش‌خورده (Clipping Fraction Paradox): جدول ۱۰ مقایسه جالب میانگین توکن‌های برش‌خورده را نشان می‌دهد. با وجود اینکه **GSPO** حدود دو مرتبه بزرگی توکن‌های بیشتری را برش می‌زند (۱۵٪ در برابر ۰.۱۳٪)، راندمان نمونه‌ای بسیار بالاتری ارائه می‌دهد؛ امری که نشان می‌دهد گرادیان‌های توکنی **GRPO** سراسر نویز مخرب بوده‌اند.

جدول ۱۰: مقایسه کسر توکن‌های برش‌خورده در طول آموزش [۴۶].

شاخص مقایسه‌ای	الگوریتم GRPO	الگوریتم GSPO
کسر توکن‌های برش‌خورده (Clipping Fraction)	۰.۰۰۱۳ (۰.۱۳٪)	**۱۵۰۰.۰ (۱۵.۰٪)**
نحوه اعمال برش	سطح تک‌توکن	سطح کل توالی
راندمان بهره‌وری نمونه (Sample Efficiency)	پایین (انباشت نویز)	بسیار بالا

ابرپارامترهای استاندارد گزارش‌شده در مقاله

جدول ۱۱ ابرپارامترهای بهینه‌سازی آموزش **GSPO** را در مقایسه با **GRPO** نشان می‌دهد:

جدول ۱۱: ابرپارامترهای استاندارد آموزش RL برای مدل **Qwen3-30B-A3B**.

ابرپارامتر	مقدار در الگوریتم GSPO
محدوده برش چپ و راست $(\varepsilon_{\text{left}}, \varepsilon_{\text{right}})$	$3 \times 10^{-4},\ 4 \times 10^{-4}$
محدوده برش در الگوریتم پایه **GRPO**	$0.20,\ 0.27$
اندازه گروه پاسخ‌ها به ازای پرسش (G)	۸
تعداد مینی‌بچ‌ها به ازای هر دسته داده‌برداری (K)	۴
نوع توزیع وزن گرادیان درون توالی	همگن با ضریب $s_i(\theta)\hat{A}_i/\lvert y_i\rvert$
نیاز به استراتژی Routing Replay	**خیر (حذف کامل)**

جمع‌بندی

الگوریتم **GSPO** با حل ناسازگاری ریاضی نمونه‌برداری اهمیت و اعمال وزن‌دهی و برش همگن در سطح کل توالی، معضل فروپاشی فاجعه‌بار آموزش را برای همیشه برطرف ساخته است. این الگوریتم با حذف نیاز به Routing Replay در مدل‌های تنک MoE و افزایش چشمگیر راندمان محاسباتی در زیرساخت‌های توزیع‌شده، به عنوان استاندارد نوین مقیاس‌پذیری یادگیری تقویتی در مدل‌های زبانی استدلالی شناخته می‌شود.

۴.۷ الگوریتم بهینه‌سازی خط‌مشی با برش مجزا و نمونه‌برداری پویا (Decoupled Clip and Dynamic Sampling Policy Optimization - DAPO)

الگوریتم **DAPO** که در سال ۲۰۲۵ توسط محققان آزمایشگاه ByteDance Seed و دانشگاه تسینگ‌هوا (Tsinghua AIR) معرفی شد [۵۳]، یک سامانه پیشرفته، مقیاس‌پذیر و کاملاً متن‌باز برای یادگیری تقویتی در مدل‌های زبانی استدلال‌گر بزرگ (Reasoning LLMs) به شمار می‌رود. این الگوریتم با برطرف کردن موانع اساسی در سناریوهای استدلال با زنجیره تفکر طولانی (Long Chain-of-Thought - Long-CoT)، موفق شد بر پایه مدل پایه **Qwen2.5-32B** [۵۴] به دقت ۵۰٪ در بنچمارک المپیاد ریاضی AIME 2024 دست یابد و با مصرف تنها ۵۰٪ از گام‌های آموزشی، از مدل پیشروی **DeepSeek-R1-Zero-Qwen-32B** [۵۵] پیشی بگیرد.

تاریخچه، انگیزه و ضرورت پیدایش الگوریتم

با ظهور پارادایم افزایش محاسبات در زمان استنتاج (Test-Time Compute Scaling) توسط مدل‌هایی همچون OpenAI o1 [۴۹] و **DeepSeek-R1** [۵۵]، استدلال مبتنی بر زنجیره تفکر طولانی به ابزار اصلی حل مسائل ریاضیات پیشرفته و برنامه‌نویسی مبدل گشت. با این وجود، به دلیل پنهان نگه‌داشتن جزئیات فنی و ترفندهای حیاتی پایدارسازی آموزش در گزارش‌های فنی این مدل‌ها، تلاش‌های مستقل جامعه متن‌باز برای بازتولید این نتایج با الگوریتم‌های رایج نظیر **GRPO** [۳۹] و **PPO** [۳۰] با ناکامی مواجه شد (دستیابی به تنها ۳۰٪ در AIME).

تحلیل‌های عمیق نشان داد که الگوریتم‌های سنتی در سناریوی استدلال‌های طولانی از چهار چالش بحرانی رنج می‌برند:

- **فروپاشی آنتروپی (Entropy Collapse):** محدودیت بازه برش استاندارد، مانع ارتقای احتمال توکن‌های نادر اما حیاتی برای اکتشاف شده و مدل را به سرعت دچار دترمینیسم زودهنگام و پاسخ‌های یکنواخت می‌کرد.
- **افت سیگنال گرادیان (Vanishing Gradient in Extreme Samples):** در پرامپت‌هایی که تمام خروجی‌های گروه درست یا همگی غلط بودند، واریانس پاداش صفر شده و گرادیان آموزشی ناپدید می‌گشت که به کاهش اندازه موثر دسته و افت شدید نسبت سیگنال به نویز (SNR) می‌انجامید.
- **ناهمگونی وزن توکن‌ها (Sample vs. Token Level Imbalance):** میانگین‌گیری در سطح نمونه در روش‌های پیشین باعث می‌شد که توکن‌های مفید در پاسخ‌های طولانی سهم گرادیان اندکی دریافت کنند و الگوهای تکراری مخرب و هذیان‌های طویل (Repetitive Gibberish) به اندازه کافی جریمه نشوند.
- **نویز ناشی از برش سخت پاسخ‌ها (Truncation Reward Noise):** جریمه کردن پاسخ‌های درستی که صرفاً به سقف طول مجاز برخورد کرده بودند، سیگنال‌های گمراه‌کننده‌ای به مدل القا می‌کرد.

الگوریتم **DAPO** برای حل جامع این معضلات، چهار نوآوری اساسی را در قالب یک چارچوب بهینه‌سازی مقیاس‌پذیر ارائه داد.

مبانی نظری و اشتقاق‌های کامل ریاضی

فرض کنید یک توالی کامل تولیدشده به صورت $\tau = (q, o_1, o_2, \ldots, o_T)$ نمایش داده شود که در آن q پرسش و o_t توکن‌های خروجی تحت خط‌مشی پارامتریزه π_θ هستند. هدف بهینه‌سازی بیشینه‌سازی

امید ریاضی پاداش است:

$$J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}[R(\tau)] = \int P(\tau;\theta)R(\tau)\,d\tau \tag{۶۸}$$

با مشتق‌گیری نسبت به بردار پارامترها θ طبق قاعده لایبنیتس و اعمال اتحاد لگاریتمی تابع امتیاز ($\nabla_\theta P = P\nabla_\theta \log P$):

$$\nabla_\theta J(\theta) = \int \nabla_\theta P(\tau;\theta)R(\tau)\,d\tau = \int P(\tau;\theta)\nabla_\theta \log P(\tau;\theta)R(\tau)\,d\tau = \mathbb{E}_{\tau\sim\pi_\theta}\left[\nabla_\theta \log P(\tau;\theta)R(\tau)\right] \tag{۶۹}$$

با توجه به ساختار خودبازگشتی مدل‌های زبانی ($P(\tau;\theta) = P(q)\prod_{t=1}^{|o|}\pi_\theta(o_t|q,o_{<t})$) و کسر خط مبنای مستقل از عمل جهت کاهش واریانس، گرادیان بر حسب تابع مزیت $\hat{A}_t$ استخراج می‌شود:

$$\nabla_\theta J(\theta) = \mathbb{E}_{q\sim\mathcal{D},o\sim\pi_\theta(\cdot|q)}\left[\sum_{t=1}^{|o|}\nabla_\theta \log \pi_\theta(o_t|q,o_{<t})\hat{A}_t\right] \tag{۷۰}$$

با استفاده از نمونه‌برداری ترجیحی (Importance Sampling) برای استفاده مجدد از داده‌های جمع‌آوری‌شده توسط خط‌مشی قدیمی $\pi_{\theta_{\text{old}}}$، نسبت احتمالاتی برای توکن t به صورت زیر تعریف می‌شود:

$$r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})} \tag{۷۱}$$

در الگوریتم **GRPO** [۳۹]، برای هر پرسش q یک گروه شامل G پاسخ $\{o_i\}_{i=1}^G$ تولید شده و مزیت هر نمونه به صورت نرمال‌سازی درون‌گروهی بدون شبکه منتقد محاسبه می‌شود:

$$\hat{A}_{i,t} = \frac{R_i - \text{mean}(\{R_i\}_{i=1}^G)}{\text{std}(\{R_i\}_{i=1}^G) + \epsilon_{\text{var}}} \tag{۷۲}$$

حذف جریمه واگرایی KL: بر خلاف روش‌های رایج RLHF که عبارت جریمه $-\beta D_{\text{KL}}(\pi_\theta \parallel \pi_{\text{ref}})$ را برای جلوگیری از انحراف مدل تحمیل می‌کنند، در استدلال‌های طولانی، توزیع مدل برای یادگیری راهبردهای جدید باید بتواند به طور چشمگیری از مدل پایه اولیه فاصله بگیرد؛ لذا در **DAPO** ترم واگرایی KL به طور کامل حذف شده است ($\beta = 0$).

فرمولاسیون ریاضی و چهار نوآوری بنیادین DAPO

تابع هدف سراسری الگوریتم **DAPO** به صورت زیر تعریف می‌شود:

$$J_{\text{DAPO}}(\theta) = \mathbb{E}_{(q,a)\sim\mathcal{D},\{o_i\}_{i=1}^G\sim\pi_{\theta_{\text{old}}}}\left[\frac{1}{\sum_{i=1}^G|o_i|}\sum_{i=1}^G\sum_{t=1}^{|o_i|}\min\left(r_{i,t}(\theta)\hat{A}_{i,t},\ \text{clip}\left(r_{i,t}(\theta),1-\varepsilon_{\text{low}},1+\varepsilon_{\text{high}}\right)\hat{A}_{i,t}\right)\right] \tag{۷۳}$$

$$\text{s.t.}\quad 0 < \sum_{i=1}^G \mathbb{I}(\text{is_equivalent}(a,o_i)) < G \tag{۷۴}$$

چهار رکن اساسی این فرمولاسیون عبارتند از:

۱. برش نامتقارن به سمت بالا (Clip-Higher): در برش متقارن کلاسیک ($\varepsilon = 0.2$)، بیشینه نسبت احتمالاتی مجاز برای توکن‌های با مزیت مثبت برابر با $1+\varepsilon = 1.2$ است. برای توکن‌های استثماری

با احتمال اولیه بالا (مثلاً $\pi_{\text{old}} = 0.9$) بیشینه احتمال مجاز 1.08 بوده که اجازه رشد تا نزدیکی ۱ را می‌دهد. اما برای توکن‌های اکتشافی نادر (مثلاً $\pi_{\text{old}} = 0.01$)، سقف رشد مجاز تنها 0.012 است ($\Delta p = +0.002$). الگوریتم **DAPO** با تفکیک آستانه‌ها و قرار دادن $\varepsilon_{\text{high}} = 0.28$ و $\varepsilon_{\text{low}} = 0.20$، فضای کافی برای ارتقای توکن‌های اکتشافی فراهم می‌سازد و مانع فروپاشی آنتروپی می‌شود.

۲. **نمونه‌برداری پویا (Dynamic Sampling):** اگر تمامی نمونه‌های گروه برای یک پرامپت درست یا همگی غلط باشند، $\hat{A}_{i,t} = 0$ شده و گرادیان آموزشی آن پرامپت صفر می‌شود. **DAPO** با اعمال شرط قید در حین گردآوری داده، نمونه‌برداری را تا پر شدن کامل بافر با دسته‌های دارای گرادیان موثر ($0 < \text{Correct} < G$) ادامه می‌دهد که این امر نسبت سیگنال به نویز گرادیان را در بالاترین سطح نگه می‌دارد.

۳. **تابع اتلاف در سطح توکن (Token-Level Policy Gradient Loss):** بر خلاف **GRPO** که ابتدا اتلاف را بر طول هر نمونه $|o_i|$ تقسیم می‌کرد، **DAPO** نرمال‌سازی را بر مجموع کل توکن‌های گروه ($\sum_{i=1}^{G} |o_i|$) انجام می‌دهد. این امر موجب می‌شود توکن‌های الگوهای استدلالی بلندمرتبه ارزش‌گذاری عادلانه شده و توکن‌های هذیان‌گونه در پاسخ‌های معیوب به شدت تنبیه شوند.

۴. **شکل‌دهی نرم پاداش طول پاسخ (Soft Overlong Punishment):** برای جلوگیری از شوک گرادیانی ناشی از برش سخت پاسخ‌ها در طول L_{max}، یک بازه حفاظتی L_{cache} تعریف شده و جریمه خطی پیوسته‌ای به پاداش درستی اضافه می‌گردد:

$$R_{\text{length}}(y) = \begin{cases} 0 & |y| \le L_{\text{max}} - L_{\text{cache}} \\ \frac{(L_{\text{max}} - L_{\text{cache}}) - |y|}{L_{\text{cache}}} & L_{\text{max}} - L_{\text{cache}} < |y| \le L_{\text{max}} \\ -1 & |y| > L_{\text{max}} \end{cases} \tag{۷۵}$$

تحلیل‌های تئوری اطلاعات و آمار ریاضی

آنتروپی شانون خط‌مشی به صورت $\mathcal{H}(\pi_\theta(\cdot|s)) = -\sum_{v \in \mathcal{V}} \pi_\theta(v|s) \log \pi_\theta(v|s)$ تعریف می‌شود. در الگوریتم‌های استاندارد، به دلیل محدودیت شدید رشد احتمالات کوچک، $\mathcal{H} \to 0$ میل کرده و توزیع احتمالاتی بر روی چند توکن محدود تیز می‌شود. راهبرد Clip-Higher باعث حفظ یک شیب ملایم مثبت در آنتروپی تولیدی در طول آموزش شده و از افت به سمت حالت دترمینستیک جلوگیری می‌نماید. همچنین نسبت سیگنال به نویز گرادیان که با $\text{SNR} = \frac{\|\mathbb{E}[g]\|_2}{\sqrt{\text{Tr}(\text{Cov}(g))}}$ تعریف می‌شود، در حضور نمونه‌برداری پویا به دلیل حذف گرادیان‌های تهی همواره در حالت ماکسیمم باقی می‌ماند.

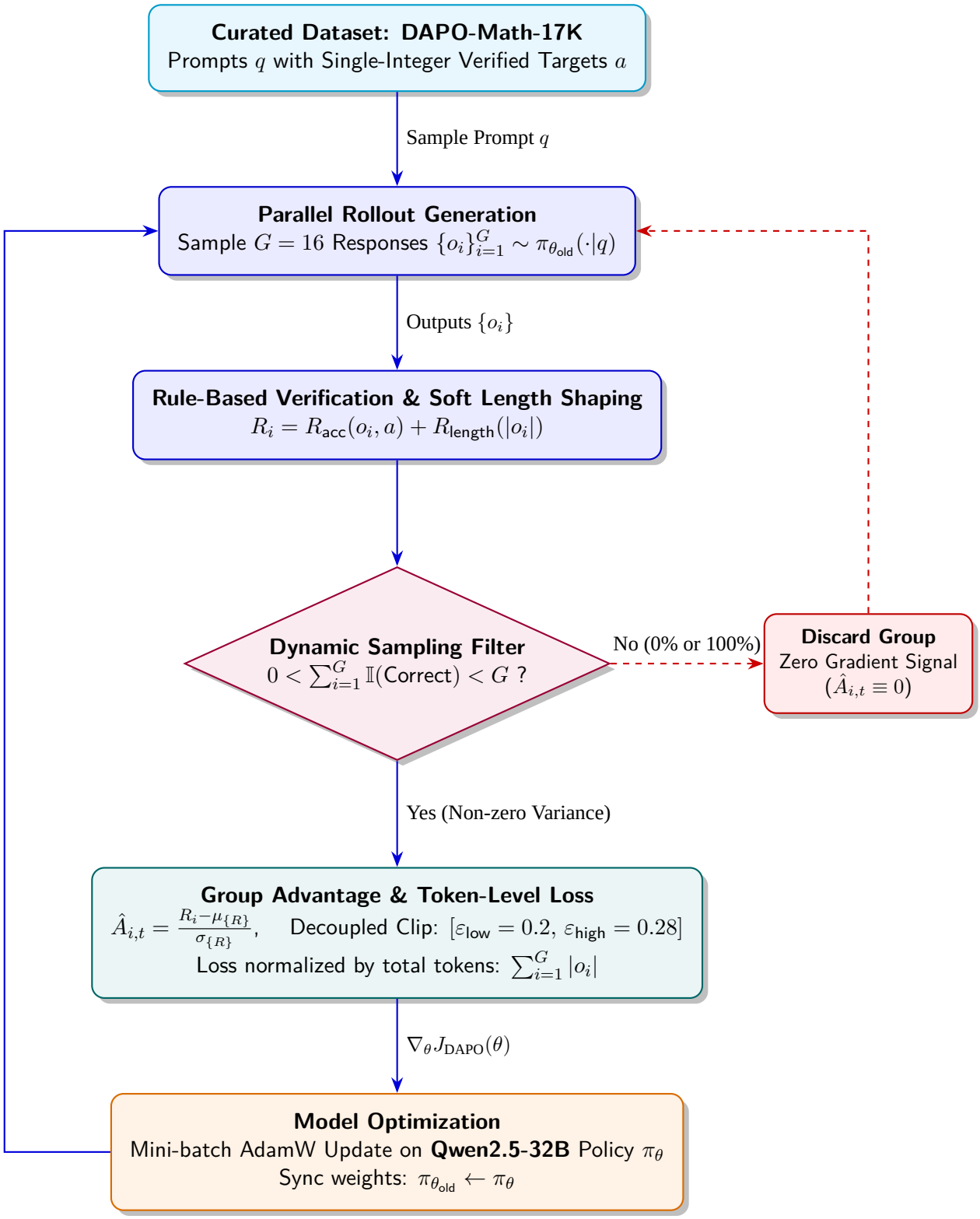
مهندسی داده و تبدیل مجموعه داده (DAPO-Math-17K)

برای بهره‌گیری از پاداش‌های قطعی مبتنی بر قاعده و حذف خطاهای ناشی از مفسرهای فرمول‌های ریاضی (Math Parsers)، مجموعه داده **DAPO-Math-17K** شامل ۱۷ هزار مسئله بازنویسی‌شده ایجاد شده است. در این فرآیند، با استفاده از مدل‌های زبانی و زنجیره تفکر، صورت مسائل به گونه‌ای بازنویسی شده‌اند که خروجی نهایی آن‌ها بدون استثنا یک **عدد صحیح منفرد** باشد؛ امری که ارزیابی پاداش را بر اساس تابع قطعی زیر مقدور می‌سازد:

$$R(\hat{y}, y) = \begin{cases} +1 & \text{is_equivalent}(\hat{y}, y) \\ -1 & \text{otherwise} \end{cases} \tag{۷۶}$$

دیاگرام معماری و جریان داده‌های الگوریتم

شکل ۴، ساختار کلی جریان داده، فیلتر نمونه‌برداری پویا، شکل‌دهی پاداش و به‌روزرسانی خط‌مشی را در الگوریتم **DAPO** نشان می‌دهد.



شکل ۴: دیاگرام جامع جریان داده، نمونه‌برداری پویا، شکل‌دهی طول و بهینه‌سازی توکن‌محور در الگوریتم **DAPO**.

شبه‌کد الگوریتم DAPO

شبه‌کد کامل مراحل اجرایی الگوریتم **DAPO** در الگوریتم ۵ ارائه شده است:

Algorithm 5 Decoupled Clip and Dynamic Sampling Policy Optimization (DAPO)

Input: Initial policy π_θ, verifiable reward function R, dataset $\mathcal{D}$, thresholds $\varepsilon_{\text{low}}, \varepsilon_{\text{high}}$, group size G, batch capacity N, update iterations μ.
for step $= 1, \ldots, M$ **do**
  Synchronize behavior policy: $\pi_{\theta_{\text{old}}} \leftarrow \pi_\theta$
  Initialize dynamic buffer $\mathcal{B}_{\text{dyn}} \leftarrow \emptyset$
  while $|\mathcal{B}_{\text{dyn}}| < N$ **do**
    Sample a prompt batch $\mathcal{D}_b \sim \mathcal{D}$
    for each question-answer pair $(q, a) \in \mathcal{D}_b$ **do**
      Sample G independent rollouts $\{o_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot|q)$
      Evaluate correctness count: $C_q = \sum_{i=1}^{G} \mathbb{I}(\text{is_equivalent}(a, o_i))$
      if $C_q == 0$ **or** $C_q == G$ **then**
        continue ▷ Dynamic Sampling: Filter out degenerate zero-variance prompts
      end if
      for each response o_i **do**
        Compute length punishment $R_{\text{length}}(|o_i|)$ via soft linear caching.
        Compute total reward $R_i = R(o_i, a) + R_{\text{length}}(|o_i|)$.
      end for
      Compute group statistics: $\mu_R = \text{mean}(\{R_i\}_{i=1}^{G})$, $\sigma_R = \text{std}(\{R_i\}_{i=1}^{G})$.
      for each response o_i **do**
        Compute advantage: $\hat{A}_{i,t} = (R_i - \mu_R)/(\sigma_R + 10^{-8})$.
        Add tuple $(q, o_i, \hat{A}_{i,t})$ to $\mathcal{B}_{\text{dyn}}$.
      end for
    end for
  end while
  for iter $= 1, \ldots, \mu$ **do**
    Compute total token count: $T_{\text{total}} = \sum_{o_i \in \mathcal{B}_{\text{dyn}}} |o_i|$.
    Compute importance ratio: $r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t}|q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|q, o_{i,<t})}$.
    Compute Token-Level Clipped Surrogate Loss:
$$J_{\text{DAPO}}(\theta) = \frac{1}{T_{\text{total}}} \sum_{i=1}^{|\mathcal{B}_{\text{dyn}}|} \sum_{t=1}^{|o_i|} \min\left(r_{i,t}(\theta)\hat{A}_{i,t},\ \text{clip}\left(r_{i,t}(\theta), 1-\varepsilon_{\text{low}}, 1+\varepsilon_{\text{high}}\right)\hat{A}_{i,t}\right).$$
    Update policy parameters via AdamW: $\theta \leftarrow \theta + \alpha \nabla_\theta J_{\text{DAPO}}(\theta)$.
  end for
end for
Output: Optimized Reasoning Policy π_θ.

نتایج تجربی، تحلیل پویایی آموزش و مطالعات موردی

اثربخشی گام‌به‌گام هر یک از تکنیک‌های پیشنهادی در جدول ۱۲ بر روی بنچمارک AIME 2024 خلاصه شده است:

پدیدار شدن رفتار خوداصلاحی (Emergence of Self-Correction): در طول آموزش با **DAPO**، رفتارهای بازبینی و خوداصلاحی عمیق بدون داده نظارت‌شده پدیدار گشت. به عنوان مثال در حل یک مسئله ترکیبیاتی، مدل پس از محاسبه مقدار کسری اشتباه $a_4 = 54.75$ برای تعداد افراد، تفکر خود را قطع کرده و بیان می‌کند:

جدول ۱۲: تحلیل ابلیشن و سهم پیشرونده تکنیک‌های **DAPO** روی مدل **Qwen2.5-32B** در بنچمارک AIME 2024 [۵۳].

مدل و راهبرد بهینه‌سازی	دقت AIME 2024 (avg@۳۲)
DeepSeek-R1-Zero-Qwen-32B (مرجع پیشین) [۵۵]	۰%.۴۷
Naive GRPO (خط مبنای استاندارد)	۰%.۳۰
+ فیلتر کردن پاسخ‌های قطع‌شده (Overlong Filtering)	۰%.۳۶
+ برش نامتقارن به سمت بالا (Clip-Higher)	۰%.۳۸
+ جریمه نرم طول پاسخ (Soft Overlong Punishment)	۰%.۴۱
+ اتلاف در سطح توکن (Token-Level Loss)	۰%.۴۲
+ نمونه‌برداری پویا (**DAPO کامل**)	**۰%.۵۰**

”Since a_4, the number of people owning all four items, must be a whole number, our current approach needs to be reconsidered...”

و سپس با بازتعریف متغیرها به پاسخ صحیح عدد صحیح دست می‌یابد.

ابرپارامترهای استاندارد گزارش‌شده در مقاله

جدول ۱۳ مقادیر دقیق ابرپارامترهای بهینه‌شده برای آموزش **DAPO** را نمایش می‌دهد:

جدول ۱۳: پیکربندی ابرپارامترهای الگوریتم **DAPO** روی مدل **Qwen2.5-32B**.

ابرپارامتر	مقدار
مدل پایه پیش‌آموزش‌دیده	**Qwen2.5-32B-Base**
فریم‌ورک محاسباتی	verl (HybridFlow) [۵۶]
بهینه‌ساز	AdamW [۵۷]
نرخ یادگیری (Learning Rate)	1×10^{-6} (ثابت با ۲۰ گام گرم‌کردن خطی)
اندازه دسته پرامپت در Rollout	۵۱۲
تعداد نمونه به ازای هر پرامپت (G)	۱۶
اندازه مینی‌دسته آموزشی	۵۱۲ (۱۶ گام به‌روزرسانی گرادیان به ازای هر Rollout)
سقف طول مجاز استدلال (L_{max})	۱۶،۳۸۴ توکن
بازه حافظه جریمه نرم (L_{cache})	۴،۰۹۶ توکن
حداکثر سقف مطلق طول تولید	۲۰،۴۸۰ توکن
حد آستانه برش پایین (ε_{low})	۲۰.۰
حد آستانه برش بالا ($\varepsilon_{\text{high}}$)	۲۸.۰
پارامترهای ارزیابی استنتاج	دما $T = 1.0$ و Top-$p = 0.7$ روی ۳۲ بار ارزیابی

جمع‌بندی

الگوریتم **DAPO** با ترکیب چهار مولفه کلیدی Clip-Higher، Dynamic Sampling، Token-Level Loss و Soft Overlong Punishment نشان داد که چالش‌های بازتولید یادگیری تقویتی در استدلال‌های طولانی ناشی از کاستی‌های ظریف در فرمول‌بندی و پویایی‌های بهینه‌سازی است. انتشار متن‌باز کدها، معماری و پایگاه داده تبدیل‌شده **DAPO-Math-17K**، زیرساختی پایدار و تکرارپذیر را برای پژوهش‌های آینده در زمینه توسعه مدل‌های هوش مصنوعی استدلال‌گر در مقیاس بزرگ فراهم آورده است.

۵.۷ الگوریتم بهینه‌سازی خط‌مشی تطبیقی نرم (Soft Adaptive Policy Optimization - SAPO)

الگوریتم **SAPO** که در اواخر سال ۲۰۲۵ توسط تیم Qwen در شرکت علی‌بابا (Alibaba Inc.) معرفی شد [۵۸]، یکی از پیشرفته‌ترین و پایدارترین روش‌های بهینه‌سازی خط‌مشی در مقیاس بزرگ برای مدل‌های زبانی استدلال‌گر (Reasoning LLMs) و مدل‌های چندوجهی (Vision-Language Models) به شمار می‌رود. این الگوریتم با بازنگری بنیادین در مکانیزم برش سخت (Hard Clipping) در روش‌های پیشین گروهی نظیر **GRPO** [۳۹] و **GSPO** [۵۹]، یک دروازه‌بندی نرم کنترل‌شده با دما (Temperature-Controlled Soft Gate) به همراه طراحی دمای نامتقارن برای گرادیان‌های مثبت و منفی معرفی می‌کند که مانع از فروپاشی زودهنگام آموزش (Early Training Collapse) شده و بهره‌وری نمونه‌ها را به حداکثر می‌رساند.

تاریخچه، انگیزه و ضرورت پیدایش الگوریتم

با گسترش آموزش یادگیری تقویتی در مدل‌های زبانی بزرگ برای انجام استدلال‌های طولانی ریاضی، برنامه‌نویسی و منطقی [۴۹، ۶۰، ۶۱]، روش‌های مبتنی بر بهینه‌سازی گروهی (Group-based Policy Optimization) به عنوان استانداردی کارآمد جایگزین روش‌های کلاسیک نظیر **PPO** شدند:

- **معضل روش‌های برش سخت در سطح توکن (GRPO):** الگوریتم **GRPO** [۳۹] با حذف مدل نقاد (Critic) و نرمال‌سازی پاداش‌ها در یک گروه از پاسخ‌ها، هزینه محاسباتی را به شدت کاهش داد. با این حال، استفاده از تابع برش ناپیوسته در سطح توکن باعث ایجاد رفتار «همه یا هیچ» شد؛ به طوری که توکن‌هایی با نسبت اهمیت خارج از بازه $[1-\varepsilon, 1+\varepsilon]$ گرادیان صفر دریافت کرده و سیگنال‌های یادگیری ارزشمند از بین می‌رفتند.
- **معضل روش‌های برش در سطح توالی (GSPO):** الگوریتم **GSPO** [۵۹] برش را به میانگین هندسی نسبت‌های کل توالی اعمال کرد تا واریانس انباشته در توالی‌های طولانی را مهار کند. اما نقطه ضعف اساسی آن شکنندگی در برابر توکن‌های پرت (Outlier Tokens) بود؛ اگر تنها چند توکن ناهمگون در یک توالی وجود داشت، کل توالی بریده شده و گرادیان تمامی توکن‌های مفید درون آن توالی مسدود می‌شد.
- **ناپایداری در مدل‌های ترکیبی از متخصصان (MoE):** در معماری‌های MoE نظیر Qwen3-30B-A3B، به دلیل ناهمگونی مسیریابی توکن‌ها (Routing Heterogeneity) به متخصصان مختلف، واریانس نسبت‌های اهمیت $r_{i,t}$ فوق‌العاده افزایش می‌یابد که این پدیده به ناپایداری‌های شدید و سقوط آموزش در الگوریتم‌های سنتی می‌انجامد.

الگوریتم **SAPO** برای حل این چالش‌ها، یک ناحیه اطمینان پیوسته، نرم و دیفرانسیل‌پذیر معرفی می‌کند که در شرایط عادی انسجام سطح توالی را حفظ کرده و در صورت وجود توکن‌های پرت، به صورت تطبیقی تنها توکن‌های مخرب را مهار می‌نماید.

مبانی نظری و اشتقاق‌های کامل ریاضی الگوریتم SAPO

یک مدل زبانی خودبازگشت به عنوان خط‌مشی تصادفی π_θ بر روی توالی توکن‌های پاسخ $y = (y_1, y_2, \ldots, y_{|y|})$ به ازای پرامپت ورودی $q \sim \mathcal{D}$ به صورت حاصل‌ضرب احتمال‌های شرطی تجزیه می‌شود:

$$\pi_\theta(y \mid q) = \prod_{t=1}^{|y|} \pi_\theta(y_t \mid q, y_{<t}) \tag{۷۷}$$

برای هر پرامپت q، گروهی متشکل از G پاسخ مستقل $\{y_1, \ldots, y_G\}$ از خط‌مشی رفتار $\pi_{\theta_{\text{old}}}$ نمونه‌برداری شده و پاداش‌های عددی $\{R_1, \ldots, R_G\}$ محاسبه می‌شوند. مقدار مزیت نرمال‌شده گروهی $(\hat{A}_i)$ برای

پاسخ i-ام برابر است با:

$$\hat{A}_{i,t} = \hat{A}_i = \frac{R_i - \text{mean}\left(\{R_j\}_{j=1}^G\right)}{\text{std}\left(\{R_j\}_{j=1}^G\right) + 10^{-8}} \tag{۷۸}$$

نسبت اهمیت در سطح توکن به صورت زیر تعریف می‌شود:

$$r_{i,t}(\theta) \triangleq \frac{\pi_\theta(y_{i,t} \mid q, y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t} \mid q, y_{i,<t})} \tag{۷۹}$$

تابع هدف بهینه‌سازی در الگوریتم SAPO:

$$J_{\text{SAPO}}(\theta) = \mathbb{E}_{q\sim\mathcal{D},\{y_i\}_{i=1}^G\sim\pi_{\theta_{\text{old}}}(\cdot|q)}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{t=1}^{|y_i|} f_{i,t}\left(r_{i,t}(\theta)\right)\hat{A}_{i,t}\right] \tag{۸۰}$$

که در آن تابع دروازه‌بندی نرم کنترل‌شده با دما $f_{i,t}(x)$ به صورت زیر تعریف می‌گردد:

$$f_{i,t}(x) = \sigma\left(\tau_{i,t}(x-1)\right)\cdot\frac{4}{\tau_{i,t}}, \qquad \tau_{i,t} = \begin{cases}\tau_{\text{pos}} & \text{if}\,\hat{A}_{i,t} > 0\\ \tau_{\text{neg}} & \text{if}\,\hat{A}_{i,t} \le 0\end{cases} \tag{۸۱}$$

در این رابطه، $\sigma(u) = \frac{1}{1+e^{-u}}$ تابع سیگموئید استاندارد است.

محاسبه تحلیلی گرادیان خط‌مشی در SAPO: با اعمال عملگر گرادیان ∇_θ به رابطه (۸۰) و استفاده از قاعده زنجیره‌ای:

$$\nabla_\theta\left[f_{i,t}(r_{i,t}(\theta))\hat{A}_{i,t}\right] = f'_{i,t}(r_{i,t}(\theta))\cdot\nabla_\theta r_{i,t}(\theta)\cdot\hat{A}_{i,t} \tag{۸۲}$$

مشتق نسبت اهمیت $\nabla_\theta r_{i,t}(\theta)$ بر اساس قاعده مشتق لگاریتمی برابر است با:

$$\nabla_\theta r_{i,t}(\theta) = \frac{\nabla_\theta \pi_\theta(y_{i,t} \mid q, y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t} \mid q, y_{i,<t})} = r_{i,t}(\theta)\nabla_\theta \log \pi_\theta(y_{i,t} \mid q, y_{i,<t}) \tag{۸۳}$$

مشتق اول تابع دروازه‌بندی $f'_{i,t}(x)$ نسبت به x:

$$f'_{i,t}(x) = \frac{4}{\tau_{i,t}}\cdot\frac{d}{dx}\left[\sigma\left(\tau_{i,t}(x-1)\right)\right] = \frac{4}{\tau_{i,t}}\cdot\tau_{i,t}\cdot\sigma'\left(\tau_{i,t}(x-1)\right) = 4\sigma'\left(\tau_{i,t}(x-1)\right) \tag{۸۴}$$

با توجه به اتحاد مشتق تابع سیگموئید $\sigma'(u) = \sigma(u)(1-\sigma(u))$، اگر $p_{i,t}(\theta) \triangleq \sigma(\tau_{i,t}(r_{i,t}(\theta)-1))$ تعریف شود:

$$w_{i,t}(\theta) \triangleq f'_{i,t}(r_{i,t}(\theta)) = 4p_{i,t}(\theta)\big(1-p_{i,t}(\theta)\big) \tag{۸۵}$$

با استفاده از تعاریف توابع هذلولوی و رابطه $\sigma(u)(1-\sigma(u)) = \frac{e^{-u}}{(1+e^{-u})^2} = \frac{1}{(e^{u/2}+e^{-u/2})^2} = \frac{1}{4\cosh^2(u/2)} = \frac{1}{4}\text{sech}^2(u/2)$، وزن گرادیان برابر می‌شود با:

$$w_{i,t}(\theta) = 4\cdot\frac{1}{4}\text{sech}^2\left(\frac{\tau_{i,t}}{2}(r_{i,t}(\theta)-1)\right) = \text{sech}^2\left(\frac{\tau_{i,t}}{2}(r_{i,t}(\theta)-1)\right) \tag{۸۶}$$

بنابراین، فرم نهایی گرادیان خط‌مشی به دست می‌آید:

$$\nabla_\theta J_{\text{SAPO}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{t=1}^{|y_i|} w_{i,t}(\theta)\, r_{i,t}(\theta)\,\nabla_\theta \log \pi_\theta(y_{i,t} \mid q, y_{i,<t})\,\hat{A}_{i,t}\right] \tag{۸۷}$$

اثبات حفظ رفتار روی خط‌مشی (On-Policy Consistency): در نقطه روی خط‌مشی که $r_{i,t}(\theta) = 1$ است:

$$p_{i,t} = \sigma(0) = \frac{1}{2} \implies w_{i,t} = 4\left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = 1 \tag{۸۸}$$

این نتیجه اثبات می‌کند که در نقطه تعادل، گرادیان نرم **SAPO** دقیقاً برابر با گرادیان خط‌مشی بدون برش $r_{i,t}\hat{A}_{i,t}\nabla_\theta \log \pi_\theta$ است و ضریب $\frac{4}{\tau_{i,t}}$ در تعریف اولیه تابع $f_{i,t}$ دقیقاً جهت تضمین این شرط مقیاس‌گذاری شده است.

تحلیل گرادیان لاجیت‌ها و اثبات ضرورت دمای نامتقارن ($\tau_{\text{neg}} > \tau_{\text{pos}}$)

برای درک علت برتری دمای نامتقارن، انتشار گرادیان در لایه خروجی لاجیت‌ها $\mathbf{z} = [z_1, \ldots, z_{|V|}]^T \in \mathbb{R}^{|V|}$ (با اندازه دایره واژگان $|V|$) را تحلیل می‌کنیم. توزیع احتمال توکن‌ها با تابع بیشینه هموار (Softmax) محاسبه می‌شود: $\pi_\theta(v) = \frac{e^{z_v}}{\sum_k e^{z_k}}$.

مشتق جزئی تابع اتلاف نسبت به لاجیت دلخواه z_v برابر است با:

$$\frac{\partial \left[\log \pi_\theta(y_{i,t})\hat{A}_{i,t}\right]}{\partial z_v} = \frac{\hat{A}_{i,t}}{\pi_\theta(y_{i,t})} \cdot \frac{\partial \pi_\theta(y_{i,t})}{\partial z_v} \tag{۸۹}$$

بر اساس ماتریس ژاکوبی تابع Softmax:

$$\frac{\partial \pi_\theta(y_{i,t})}{\partial z_{y_{i,t}}} = \pi_\theta(y_{i,t})\big(1 - \pi_\theta(y_{i,t})\big) \quad \text{(توکن نمونه‌برداری‌شده):} \; v = y_{i,t} \text{ اگر} \tag{۹۰}$$

$$\frac{\partial \pi_\theta(y_{i,t})}{\partial z_v} = -\pi_\theta(y_{i,t})\pi_\theta(v) \quad \text{(سایر توکن‌ها):} \; v \neq y_{i,t} \text{ اگر} \tag{۹۱}$$

با جایگذاری روابط فوق در رابطه (۸۹)، گرادیان لاجیت‌ها حاصل می‌شود:

$$\frac{\partial \mathcal{L}_{i,t}}{\partial z_v} = \begin{cases} \big(1 - \pi_\theta(y_{i,t})\big)\hat{A}_{i,t} & \text{if} v = y_{i,t} \quad \text{(Sampled Token)} \\ -\pi_\theta(v)\hat{A}_{i,t} & \text{if} v \neq y_{i,t} \quad \text{(Unsampled Tokens)} \end{cases} \tag{۹۲}$$

تحلیل تئوریک رفتار نامتقارن:

- **در شرایط مزیت مثبت ($\hat{A}_{i,t} > 0$):** لاجیت توکن نمونه‌برداری‌شده افزایش یافته و لاجیت تمامی توکن‌های دیگر با ضریب $-\pi_\theta(v)\hat{A}_{i,t} < 0$ به آرامی کاهش می‌یابد. این فرآیند ذاتاً همگرا و متمرکز است.
- **در شرایط مزیت منفی ($\hat{A}_{i,t} < 0$):** لاجیت توکن نمونه‌برداری‌شده کاهش می‌یابد؛ اما برای تمامی $|V| - 1$ توکن دیگر، مقدار $-\pi_\theta(v)\hat{A}_{i,t} > 0$ مثبت می‌شود! یعنی لاجیت صدها هزار توکن نامربوط به طور همزمان افزایش می‌یابد. این پدیده باعث پخش نویز گرادیانی شدید در دایره واژگان بزرگ شده و عامل اصلی ناپایداری آموزش است.

SAPO با تنظیم $\tau_{\text{neg}} > \tau_{\text{pos}}$، نرخ تضعیف گرادیان برای نمونه‌های با مزیت منفی را سریع‌تر می‌کند و بدین ترتیب نویز مخرب ناشی از انتشار مزیت منفی را مهار می‌نماید.

چارچوب یکپارچه سوروگیت و اثبات همگرایی به دروازه توالی (اتصال SAPO–GSPO)

فرض‌های تحلیلی:

- **فرض A۱ (گام‌های کوچک و نزدیکی به خط‌مشی):** $r_{i,t}(\theta) \approx 1 \implies \log r_{i,t}(\theta) \approx r_{i,t}(\theta) - 1$.
- **فرض A۲ (پراکندگی پایین درون توالی):** با تعریف $z_{i,t} \triangleq \log r_{i,t}(\theta)$ و میانگین $\mu_i \triangleq \frac{1}{|y_i|}\sum_t z_{i,t} = \log s_i(\theta)$، واریانس درون‌توالی $\mathsf{Var}_i(\theta) \triangleq \frac{1}{|y_i|}\sum_t (z_{i,t} - \mu_i)^2$ برای اکثر توالی‌ها مقداری بسیار کوچک است.

تحت فرض A۱ داریم $f'_{i,t}(r_{i,t}) = \mathsf{sech}^2\left(\frac{\tau_i}{2}(r_{i,t} - 1)\right) \approx \mathsf{sech}^2\left(\frac{\tau_i}{2}\log r_{i,t}\right) = g_{\tau_i}(z_{i,t})$ که در آن $g_\tau(z) \triangleq \mathsf{sech}^2(\frac{\tau}{2}z)$ است.

با نوشتن بسط تیلور مرتبه دوم تابع $g_{\tau_i}(z_{i,t})$ حول نقطه میانگین $\mu_i = \log s_i(\theta)$:

$$g_{\tau_i}(z_{i,t}) = g_{\tau_i}(\mu_i) + g'_{\tau_i}(\mu_i)(z_{i,t} - \mu_i) + \frac{1}{2}g''_{\tau_i}(\xi_{i,t})(z_{i,t} - \mu_i)^2 \tag{۹۳}$$

با میانگین‌گیری روی تمام توکن‌های توالی، جمله خطی اول دقیقاً صفر می‌شود:

$$\frac{1}{|y_i|}\sum_{t=1}^{|y_i|} g_{\tau_i}(z_{i,t}) = g_{\tau_i}(\mu_i) + \frac{1}{2|y_i|}\sum_{t=1}^{|y_i|} g''_{\tau_i}(\xi_{i,t})(z_{i,t} - \mu_i)^2 \tag{۹۴}$$

با محاسبه مشتق دوم $g_\tau(z) = \mathsf{sech}^2(\alpha z)$ با $\alpha = \frac{\tau}{2}$:

$$g''_\tau(z) = \alpha^2\big(4\mathsf{sech}^2(\alpha z) - 6\mathsf{sech}^4(\alpha z)\big) \implies \sup_z |g''_\tau(z)| = 2\alpha^2 = \frac{\tau^2}{2} \tag{۹۵}$$

بنابراین، کران یکنواخت اختلاف میانگین دروازه توکن‌ها و دروازه توالی $D_i(\theta)$ حاصل می‌شود:

$$D_i(\theta) \triangleq \left|\frac{1}{|y_i|}\sum_{t=1}^{|y_i|} g_{\tau_i}(z_{i,t}) - g_{\tau_i}\big(\log s_i(\theta)\big)\right| \le \frac{1}{2}\sup_z |g''_{\tau_i}(z)|\mathsf{Var}_i(\theta) = \frac{\tau_i^2}{4}\mathsf{Var}_i(\theta) \tag{۹۶}$$

از آنجا که $D_i(\theta) \approx 0$ است، گرادیان **SAPO** به فرم سطح توالی تقلیل می‌یابد:

$$\nabla_\theta J_{\mathsf{SAPO}}(\theta) \approx \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G} \mathsf{sech}^2\left(\frac{\tau_i}{2}\log s_i(\theta)\right)\nabla_\theta \log s_i(\theta)\hat{A}_i\right] \tag{۹۷}$$

این اثبات نشان می‌دهد که **SAPO** در حالت عادی رفتاری کاملاً منسجم و هموار در سطح توالی (مشابه **GSPO**) دارد، اما در برخورد با توکن‌های پرت، به جای قطع کل گرادیان توالی، تنها توکن خطاکار را تضعیف می‌کند.

تحلیل‌های تئوری اطلاعات، آمار و دینامیک‌های ناحیه اطمینان

بر اساس بسط تیلور مرتبه دوم واگرایی کولبک-لایبلر موضعی در سطح توکن:

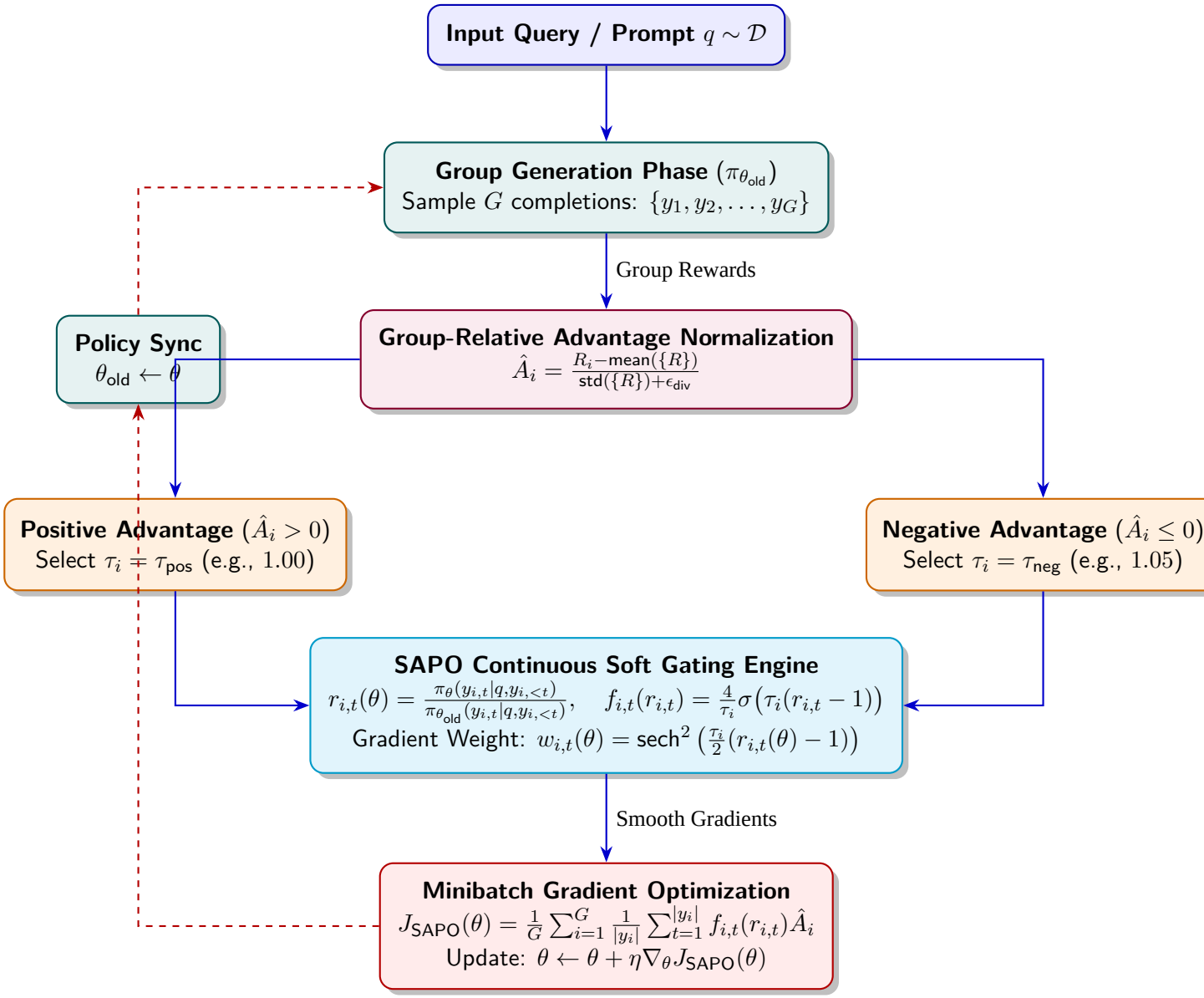
$$D_{\mathsf{KL}}(\pi_{\theta_{\mathsf{old}}} \parallel \pi_\theta) \approx \frac{1}{2}\mathbb{E}\left[(\log r_{i,t}(\theta))^2\right] \tag{۹۸}$$

در روش‌های برش سخت (**GRPO** و **GSPO**)، تابع جریمه در مرزهای برش دارای ناپیوستگی در مشتق اول است که شوک‌های بهینه‌سازی ایجاد می‌کند. در مقابل، تابع دروازه‌بندی **SAPO** از رده توابع بی‌نهایت بار مشتق‌پذیر (C^∞) بوده و یک ناحیه اطمینان پیوسته را شکل می‌دهد.

بررسی‌های تجربی روی بیش از 10^5 توالی و 10^9 توکن نشان می‌دهد که توزیع واریانس درون‌توالی $\mathsf{Var}_i(\theta)$ در مدل‌های متراکم (Dense) نظیر Qwen3-4B کمتر از 0.005 و در مدل‌های ترکیبی از متخصصان (MoE) نظیر Qwen3-30B-A3B کمتر از 0.020 است. در هر دو حالت کران خطا $D_i(\theta) \le \frac{\tau_i^2}{4}\mathsf{Var}_i(\theta)$ زیر 0.005 باقی مانده و تقلیل تئوریک را به طور کامل در عمل پشتیبانی می‌کند.

دیاگرام معماری و جریان داده‌های الگوریتم

شکل ۵ ساختار پردازش داده، ارزیابی گروهی، اعمال دمای نامتقارن، دروازه‌بندی نرم و به‌روزرسانی پارامترها در الگوریتم **SAPO** را نمایش می‌دهد.



شکل ۵: دیاگرام جامع جریان داده، انتخاب دمای نامتقارن، دروازه‌بندی نرم و بهینه‌سازی خطمشی در الگوریتم **SAPO**.

شبه‌کد الگوریتم SAPO

شبه‌کد استاندارد پیاده‌سازی بهینه‌سازی خطمشی تطبیقی نرم در الگوریتم ۶ ارائه شده است.

Algorithm 6 Soft Adaptive Policy Optimization (SAPO)

Input: Policy parameters θ_0, behavior policy checkpoint $\theta_{\text{old}} = \theta_0$, prompt dataset $\mathcal{D}$, group size G, learning rate η, positive temperature τ_{pos}, negative temperature τ_{neg}, number of mini-batches M, epochs K.
for iteration $= 1, 2, \ldots, N_{\text{iters}}$ **do**
  Sample a batch of queries $\{q_1, q_2, \ldots, q_B\} \sim \mathcal{D}$.
  Initialize rollout storage buffer $\mathcal{B}_{\text{rollout}} \leftarrow \emptyset$.
  for each query $q \in \{q_1, \ldots, q_B\}$ **do**
    Sample G responses $\{y_1, \ldots, y_G\} \sim \pi_{\theta_{\text{old}}}(\cdot \mid q)$.
    Compute rewards $\{R_1, \ldots, R_G\}$ via task-specific environment/verifier.
    Compute mean $\mu_R = \frac{1}{G}\sum_{j=1}^{G} R_j$ and std $\sigma_R = \sqrt{\frac{1}{G}\sum_{j=1}^{G}(R_j - \mu_R)^2} + 10^{-8}$.
    for $i = 1, \ldots, G$ **do**
      Compute group-normalized advantage: $\hat{A}_i = (R_i - \mu_R)/\sigma_R$.
      Store cached old log-probabilities $\{\log \pi_{\theta_{\text{old}}}(y_{i,t} \mid q, y_{i,<t})\}_{t=1}^{|y_i|}$.
      Append tuple $(q, y_i, \hat{A}_i, \{\log \pi_{\theta_{\text{old}}}(y_{i,t})\}_{t=1}^{|y_i|})$ to $\mathcal{B}_{\text{rollout}}$.
    end for
  end for
  Partition $\mathcal{B}_{\text{rollout}}$ into M disjoint mini-batches $\{\mathcal{M}_1, \ldots, \mathcal{M}_M\}$.
  for epoch $= 1, \ldots, K$ **do**
    for each mini-batch $\mathcal{M}_b$ **do**
      Initialize mini-batch loss: $\mathcal{L}_{\text{mini}}(\theta) \leftarrow 0$.
      for each sample $(q, y_i, \hat{A}_i, \text{log-probs}) \in \mathcal{M}_b$ **do**
        Select asymmetric temperature: $\tau_i \leftarrow \tau_{\text{pos}}$ if $\hat{A}_i > 0$ else τ_{neg}.
        Compute current log-probabilities $\{\log \pi_\theta(y_{i,t} \mid q, y_{i,<t})\}_{t=1}^{|y_i|}$.
        Initialize sequence objective: $\mathcal{L}_{\text{seq}} \leftarrow 0$.
        for $t = 1, \ldots, |y_i|$ **do**
          Compute token ratio: $r_{i,t}(\theta) = \exp\left(\log \pi_\theta(y_{i,t}) - \log \pi_{\theta_{\text{old}}}(y_{i,t})\right)$.
          Compute soft gate: $f_{i,t}(r_{i,t}) = \frac{4}{\tau_i} \cdot \sigma\left(\tau_i(r_{i,t}(\theta) - 1)\right)$.
          $\mathcal{L}_{\text{seq}} \leftarrow \mathcal{L}_{\text{seq}} + f_{i,t}(r_{i,t}(\theta)) \cdot \hat{A}_i$.
        end for
        $\mathcal{L}_{\text{mini}}(\theta) \leftarrow \mathcal{L}_{\text{mini}}(\theta) - \frac{1}{|y_i|}\mathcal{L}_{\text{seq}}$.
      end for
      Update parameters: $\theta \leftarrow \theta - \eta \cdot \nabla_\theta \left(\frac{1}{|\mathcal{M}_b|}\mathcal{L}_{\text{mini}}(\theta)\right)$.
    end for
  end for
  Synchronize behavior policy checkpoint: $\theta_{\text{old}} \leftarrow \theta$.
end for

نتایج تجربی و ارزیابی بنچمارک‌ها

عملکرد الگوریتم **SAPO** در مقایسه با **GRPO-R2** و **GSPO** روی مدل استدلال‌گر Qwen3-30B-A3B-Base در جدول ۱۴ خلاصه شده است.

تحلیل ابلیشن دمای نامتقارن (τ_{neg} در برابر τ_{pos}): جدول ۱۵ تأثیر تنظیم دماهای مختلف را نشان می‌دهد. بر اساس این نتایج، تعیین دمای بالاتر برای توکن‌های با مزیت منفی ($\tau_{\text{neg}} > \tau_{\text{pos}}$) نقش حیاتی در جلوگیری از انحرافات گرادیانی و تثبیت آموزش ایفا می‌کند.

آموزش مدل‌های چندوجهی (Qwen3-VL): در ارزیابی‌های مقیاس بزرگ روی سری مدل‌های چندوجهی Qwen3-VL-30B-A3B در چهار بنچمارک چندوجهی و متنی شامل AIME25 (با ۳۲ نمونه)، Live-

جدول ۱۴: مقایسه عملکرد نهایی و پایداری در بنچمارک‌های استدلال ریاضی (Pass@1 روی ۱۶ نمونه) [۵۸].

الگوریتم	AIME 2025	HMMT 2025	BeyondAIME	وضعیت پایداری آموزش
GRPO-R2 [۳۹، ۵۹]	۵۶.۷۶٪	۵۷.۰٪	۵۱.۵٪	سقوط در گام ۷۵۰ (Collapse)
GSPO [۵۹]	۷۹.۰٪	۶۱.۰٪	۵۴.۰٪	سقوط در گام ۱۲۵۰ (Collapse)
SAPO (پیشنهادی)	**۸۲.۲٪**	**۶۴.۵٪**	**۵۸.۸٪**	**کاملاً پایدار (> 1750 گام)**

جدول ۱۵: تحلیل ابلیشن تنظیمات دما در الگوریتم **SAPO** روی مدل Qwen3-30B-A3B [۵۸].

پیکربندی دما	مقادیر	دقت در AIME25	دینامیک آموزش
$\tau_{\text{neg}} < \tau_{\text{pos}}$	$\tau_{\text{neg}} = 0.95,\ \tau_{\text{pos}} = 1.00$	۷۱.۰٪	ناپایداری بسیار شدید و سقوط زودهنگام
$\tau_{\text{neg}} = \tau_{\text{pos}}$	$\tau_{\text{neg}} = 1.00,\ \tau_{\text{pos}} = 1.00$	۷۸.۵٪	پایداری متوسط با افت‌های مقطعی
$\tau_{\text{neg}} > \tau_{\text{pos}}$ (اصلی)	$\tau_{\text{neg}} = \mathbf{1.05},\ \tau_{\text{pos}} = \mathbf{1.00}$	**۸۲.۲٪**	**بالاترین پایداری و رشد پیوسته**

CodeBench v6 [۶۲] (با ۸ نمونه)، ZebraLogic [۶۳] و MathVision [۶۴]، الگوریتم **SAPO** تحت بودجه محاسباتی یکسان موفق شد به امتیاز اعتبارسنجی تجمیعی **۷۶.۰** دست یابد و با اختلاف معنادار از **GSPO** (با امتیاز ۷۴۵.۰) و **GRPO-R2** (با امتیاز ۷۳۸.۰) پیشی بگیرد.

جمع‌بندی

الگوریتم **SAPO** با جایگزین کردن برش سخت ناپیوسته با یک ناحیه اطمینان نرم (sech^2) و اعمال مکانیزم تنظیم دمای نامتقارن برای مهار نویز گرادیان‌های منفی، پایداری بهینه‌سازی خط‌مشی را متحول ساخته است. این الگوریتم بدون نیاز به سازوکارهای پرهزینه نظیر بازپخش مسیریابی (Routing Replay) در مدل‌های MoE، بالاترین دقت و بهره‌وری داده را به نمایش گذاشته و به عنوان چارچوبی استاندارد برای آموزش مدل‌های پیشرفته نسل جدید زبانی و چندوجهی شناخته می‌شود.

۶.۷ الگوریتم بهینه‌سازی خط‌مشی با ضریب اهمیت برش‌خورده (Clipped IS-weight Policy Optimization - CISPO)

الگوریتم **CISPO** یک الگوریتم پیشرفته در حوزه یادگیری تقویتی برون‌پویا (Off-Policy Reinforcement Learning) است که توسط آزمایشگاه MiniMax در سال ۲۰۲۵ و در جریان توسعه مدل استدلال‌گر بزرگ **MiniMax-M1** معرفی شد [۵۱]. این الگوریتم با حذف کامل قید سنتی ناحیه اطمینان (Trust Region) روی گرادیان توکن‌ها و در عوض اعمال برش (Clipping) صرفاً بر روی وزن‌های نمونه‌برداری اهمیت (Importance Sampling Weights)، انقلابی در مقیاس‌پذیری پایدار و کارآمد فرآیند تفکر طولانی‌مدت (Long Chain-of-Thought) پدید آورده است.

تاریخچه، انگیزه و ضرورت پیدایش الگوریتم

با پیدایش مدل‌های استدلال‌گر نظیر **OpenAI o1** و **DeepSeek-R1** [۵۰]، مقیاس‌پذیری محاسبات زمان آزمون (Test-Time Compute Scaling) به یکی از ارکان اصلی پیشرفت مدل‌های زبانی بدل شد. در این پارادایم، طول استدلال مدل به ده‌ها هزار توکن افزایش می‌یابد تا راه‌حل مسائل پیچیده المپیادی، ریاضیات و مهندسی نرم‌افزار کشف شود [۶۵].

با این حال، الگوریتم‌های استاندارد یادگیری تقویتی مانند **PPO** [۳۰] و **GRPO** [۳۹] در این زمینه با یک نقیصه ساختاری بنیادین تحت عنوان **«پاتولوژی برش توکن» (The Pathology of Token Clipping)** مواجه می‌شوند:

- **نقش حیاتی توکن‌های بازتابی (Reflective Tokens):** رفتارهای کلیدی در استدلال زنجیره‌ای طولانی مانند تفکر انتقادی و تصحیح مسیر، به توکن‌های کلیدی و اتصال‌دهنده‌ای نظیر «However»، «Wait»، «Recheck» و «Aha» وابسته‌اند. این توکن‌ها نقش دوشاخه‌شدن (Forks) در گراف مسیرهای استدلال را ایفا می‌کنند.
- **احتمال پایه پایین و جهش نسبت اهمیت:** این توکن‌ها در مدل پایه دارای احتمال اولیه اندکی هستند ($\pi_{\theta_{\text{old}}} \ll 1$). هنگامی که مدل پاداش مثبت دریافت می‌کند، در اولین به‌روزرسانی‌ها احتمال این توکن‌ها افزایش سریعی پیدا کرده و در نتیجه نسبت اهمیت $r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t})}{\pi_{\theta_{\text{old}}}(o_{i,t})}$ به سرعت از حد مجاز $(1+\epsilon)$ فراتر می‌رود.
- **صفر شدن گرادیان در PPO و GRPO:** در الگوریتم‌های سنتی، به محض عبور نسبت اهمیت از سقف $1+\epsilon$، گرادیان مربوط به آن توکن صفر می‌شود. در سناریوهای بهینه‌سازی برون‌پویا که شامل چندین دور به‌روزرسانی ریزدسته (مثلاً ۱۶ دور در هر بچ داده) هستند، این توکن‌های حیاتی پس از اولین دور کاملاً از فرآیند یادگیری کنار گذاشته می‌شوند.
- **تخریب پویایی آنتروپی و مهار CoT:** حذف گرادیان توکن‌های بازتابی باعث سقوط آنتروپی [۶۶، ۶۷]، ناپایداری بهینه‌سازی و مهار رفتارهای استدلالی بلندمدت می‌شود.

اگرچه الگوریتم **DAPO** [۵۳] تلاش نمود با بالا بردن سقف برش بالایی این چالش را کاهش دهد، اما این راهکار در چرخه‌های آف‌پالیسی طولانی ناکافی بوده و با کندی همگرایی همراه است. **CISPO** با بازطراحی ساختار تابع هدف، این مشکل را به‌صورت پایه‌ای مرتفع می‌سازد.

مبانی نظری و اشتقاق‌های کامل ریاضی از اصول اولیه

مسئله تصمیم‌گیری زبانی را در قالب یک فرآیند تصمیم‌گیری مارکوف با توزیع پرسش‌های $q \sim \mathcal{D}$ و توالی پاسخ $o = (o_1, o_2, \ldots, o_{|o|})$ مدل‌سازی می‌کنیم. توزیع احتمال خودهمبسته تولید پاسخ تحت خط‌مشی پارامتری π_θ عبارت است از:

$$P(o \mid q; \theta) = \prod_{t=1}^{|o|} \pi_\theta(o_t \mid q, o_{<t}) \tag{۹۹}$$

تابع هدف امید ریاضی پاداش در خط‌مشی به صورت زیر تعریف می‌شود:

$$J(\theta) = \mathbb{E}_{q\sim\mathcal{D}, o\sim\pi_\theta}[R(q,o)] = \sum_q P(q) \int_{\Omega_o} P(o \mid q; \theta) R(q,o)\, do \tag{۱۰۰}$$

با مشتق‌گیری نسبت به بردار پارامترهای $\theta \in \mathbb{R}^d$ و بهره‌گیری از اتحاد لگاریتمی ($\nabla_\theta P = P \nabla_\theta \log P$):

$$\begin{aligned}
\nabla_\theta J(\theta) &= \sum_q P(q) \int_{\Omega_o} \nabla_\theta P(o \mid q; \theta) R(q,o)\, do \\
&= \sum_q P(q) \int_{\Omega_o} P(o \mid q; \theta) \nabla_\theta \log P(o \mid q; \theta) R(q,o)\, do \\
&= \mathbb{E}_{q\sim\mathcal{D}, o\sim\pi_\theta}\left[\nabla_\theta \log P(o \mid q; \theta) R(q,o)\right]
\end{aligned} \tag{۱۰۱}$$

با اعمال قاعده زنجیره‌ای بر لگاریتم احتمال زنجیره توکن‌ها:

$$\nabla_\theta \log P(o \mid q; \theta) = \sum_{t=1}^{|o|} \nabla_\theta \log \pi_\theta(o_t \mid q, o_{<t}) \tag{۱۰۲}$$

در بهینه‌سازی برون‌پویا (Off-Policy) که نمونه‌ها از خط‌مشی جمع‌آوری‌کننده قدیمی $\pi_{\theta_{\text{old}}}$ نمونه‌برداری می‌شوند، با استفاده از قضیه نمونه‌برداری اهمیت (Importance Sampling Theorem) داریم:

$$\mathbb{E}_{x\sim p}[f(x)] = \int p(x)f(x)\,dx = \int q(x)\frac{p(x)}{q(x)}f(x)\,dx = \mathbb{E}_{x\sim q}\left[\frac{p(x)}{q(x)}f(x)\right] \tag{۱۰۳}$$

نسبت وزن اهمیت توکنی برای توکن t-ام از پاسخ i-ام به صورت زیر تعریف می‌گردد:

$$r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})} \tag{۱۰۴}$$

در فرمولاسیون **GRPO** [۳۹]، برای هر پرسش q تعداد G خروجی $\{o_i\}_{i=1}^G$ تولید شده و مزیت درون‌گروهی نرمال‌شده بدون نیاز به مدل نقاد (Critic-Free) محاسبه می‌شود:

$$\hat{A}_{i,t} = \hat{A}_i = \frac{R_i - \text{mean}(\{R_j\}_{j=1}^G)}{\text{std}(\{R_j\}_{j=1}^G) + \delta} \tag{۱۰۵}$$

تابع هدف REINFORCE با تصحیح توزیع برای به‌روزرسانی‌های آفلاین با بهره‌گیری از عملگر توقف مشتق (Stop-Gradient) $\text{sg}(\cdot)$ به صورت زیر بیان می‌شود:

$$\mathcal{J}_{\text{REINFORCE}}(\theta) = \mathbb{E}_{(q,a)\sim\mathcal{D}, o_i\sim\pi_{\theta_{\text{old}}}}\left[\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\text{sg}(r_{i,t}(\theta))\hat{A}_{i,t}\log\pi_\theta(o_{i,t} \mid q, o_{i,<t})\right] \tag{۱۰۶}$$

فرمولاسیون اختصاصی و مشتق گرادیان CISPO

به جای اعمال برش بر روی کل تابع هدف به‌روزرسانی که موجب قطع گرادیان می‌شود، الگوریتم **CISPO** برش را مستقیماً بر روی ضریب اهمیت اعمال کرده و آن را در داخل عملگر توقف مشتق قرار می‌دهد:

$$\hat{r}_{i,t}(\theta) = \text{clip}\left(r_{i,t}(\theta),\ 1-\epsilon_{\text{low}}^{IS},\ 1+\epsilon_{\text{high}}^{IS}\right) \tag{۱۰۷}$$

تابع هدف نهایی الگوریتم **CISPO** با تلفات در سطح توکن (Token-Level Loss) عبارت است از:

$$\mathcal{J}_{\text{CISPO}}(\theta) = \mathbb{E}_{q\sim\mathcal{D},\{o_i\}_{i=1}^G\sim\pi_{\theta_{\text{old}}}}\left[\frac{1}{\sum_{i=1}^G|o_i|}\sum_{i=1}^G\sum_{t=1}^{|o_i|}\text{sg}\left(\hat{r}_{i,t}(\theta)\right)\hat{A}_{i,t}\log\pi_\theta(o_{i,t} \mid q, o_{i,<t})\right] \tag{۱۰۸}$$

با محاسبه صریح گرادیان نسبت به پارامترهای θ:

$$\nabla_\theta\mathcal{J}_{\text{CISPO}}(\theta) = \mathbb{E}\left[\frac{1}{\sum_{i=1}^G|o_i|}\sum_{i=1}^G\sum_{t=1}^{|o_i|}\hat{r}_{i,t}(\theta)\hat{A}_{i,t}\nabla_\theta\log\pi_\theta(o_{i,t} \mid q, o_{i,<t})\right] \tag{۱۰۹}$$

تمایز بنیادین گرادیان: در معادله فوق، از آنجا که $\hat{r}_{i,t}(\theta) \geq 1-\epsilon_{\text{low}}^{IS} > 0$ همواره مقداری اکیداً مثبت است، مقدار مشتق $\nabla_\theta\log\pi_\theta$ **هرگز صفر نمی‌شود**. در نتیجه تمامی توکن‌ها با ضریب مقیاس متناسب و کران‌دار در هدایت گرادیان شبکه مشارکت خواهند داشت. در کاربرد عملی، کران پایین حذف شده ($\epsilon_{\text{low}}^{IS} \to \infty$) و تنها $\epsilon_{\text{high}}^{IS}$ تنظیم می‌شود.

فرمولاسیون تعمیم‌یافته و مقایسه‌ای (A Unified Formulation)

برای مقایسه روش‌های برش، می‌توان یک فرمولاسیون جامع با معرفی ماسک توکنی $M_{i,t}$ تعریف نمود:

$$\mathcal{J}_{\text{unify}}(\theta) = \mathbb{E}\left[\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\text{sg}\left(\hat{r}_{i,t}(\theta)\right)\hat{A}_{i,t}\log\pi_\theta(o_{i,t}\mid q, o_{i,<t})M_{i,t}\right] \tag{۱۱۰}$$

که در آن ماسک ناحیه اطمینان برابر است با:

$$M_{i,t} = \begin{cases} 0 & \text{if}\,\hat{A}_{i,t} > 0 \text{ and } r_{i,t}(\theta) > 1+\epsilon_{\text{high}} \\ 0 & \text{if}\,\hat{A}_{i,t} < 0 \text{ and } r_{i,t}(\theta) < 1-\epsilon_{\text{low}} \\ 1 & \text{otherwise} \end{cases} \tag{۱۱۱}$$

جدول ۱۶ تفاوت ساختاری این الگوریتم‌ها را به تصویر می‌کشد.

جدول ۱۶: مقایسه ساختاری الگوریتم‌های یادگیری تقویتی برون‌پویا بر پایه فرمولاسیون یکپارچه.

الگوریتم	تعریف وزن اهمیت $\hat{r}_{i,t}$	وضعیت ماسک $M_{i,t}$	حفظ گرادیان توکن‌های نادر	نیاز به مدل نقاد (Critic)
PPO [۳۰]	$r_{i,t}(\theta)$	شرطی (حذف گرادیان در خروج از مرز)	خیر	بله
GRPO [۳۹]	$r_{i,t}(\theta)$	شرطی (حذف گرادیان در خروج از مرز)	خیر	خیر
DAPO [۵۳]	$r_{i,t}(\theta)$	شرطی با ϵ_{high} بسیار بزرگتر	تا حدی	خیر
CISPO [۵۱]	$\text{clip}(r_{i,t}, 1-\epsilon_{\text{low}}^{IS}, 1+\epsilon_{\text{high}}^{IS})$	**همواره فعال ($M_{i,t}=1$)**	**کامل (۱۰۰٪ توکن‌ها)**	**خیر**

تحلیل‌های تئوری اطلاعات، آمار و پویایی آنتروپی

- **توازن بایاس و واریانس (Bias-Variance Tradeoff):** واریانس برآوردگر نمونه‌برداری اهمیت متناسب با واگرایی رنی مرتبه دوم $d_2(\pi_\theta \parallel \pi_{\theta_{\text{old}}})$ رشد می‌کند. اعمال برش مستقیم بر وزن $\hat{r}_{i,t} \leq 1+\epsilon_{\text{high}}^{IS}$، کران بالای واریانس گرادیان را محبوس می‌سازد:

$$\text{Var}\left[\nabla_\theta \mathcal{J}_{\text{CISPO}}\right] \leq (1+\epsilon_{\text{high}}^{IS})^2 \cdot \text{Var}\left[\hat{A}_{i,t}\nabla_\theta \log \pi_\theta\right] \tag{۱۱۲}$$

این رویکرد مقداری بایاس اندک وارد می‌کند اما مانع از واگرایی واریانس در دنباله‌های طولانی ($L \geq 40$K) می‌شود.

- **تثبیت آنتروپی شانون و عدم نیاز به جریمه KL:** آنتروپی در سطح توکن به صورت $\mathcal{H}(\pi_\theta) = -\sum_v \pi_\theta(v)\log\pi_\theta(v)$ محاسبه می‌شود. از آنجا که گرادیان توکن‌های اقلیت با آنتروپی بالا حذف نمی‌شود، توزیع مدل دچار فروپاشی نابهنگام (Premature Collapse) نشده و نیاز به افزودن جریمه مستقیم واگرایی کولبک-لایبلر (D_{KL}) منتفی می‌گردد [۶۷].

چالش‌های مقیاس‌پذیری در معماری‌های هیبرید و راهکارهای مهندسی

مدل **MiniMax-M1** از ساختار ترکیبی متشکل از ۷ بلوک Transnormer با مکانیزم توجه خطی صاعقه‌ای (Lightning Attention) [۶۸] به ازای هر ۱ بلوک توجه استاندارد Softmax بهره می‌برد. آموزش یادگیری تقویتی در مقیاس ۴۵۶ میلیارد پارامتر با این معماری نیازمند راهکارهای مهندسی زیر بود:

۱. **رفع عدم تطابق دقت محاسباتی (Precision Mismatch):** تفاوت جزئی در خروجی هسته‌های محاسباتی آموزش و استنتاج مانع رشد پاداش می‌شد. با ارتقای دقت لایه خروجی سر پیش‌بینی کلمات (LM Output Head) به FP32، همبستگی پیرسون احتمالات از ۹۸.۰ به بالای ۹۹۷.۰ افزایش یافت.

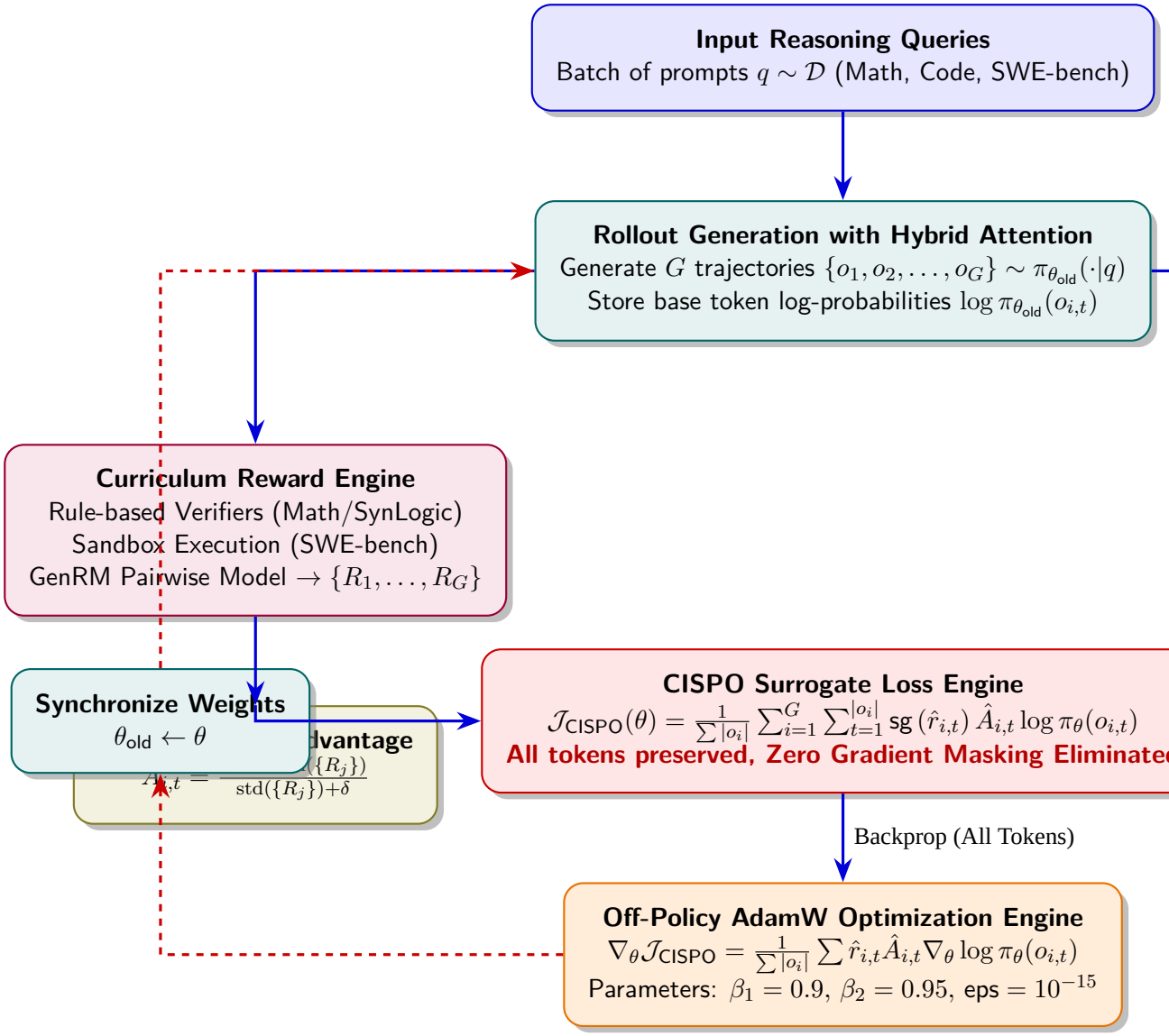
۲. **پیکربندی بهینه‌ساز AdamW:** با توجه به بازه گسترده بزرگی گرادیان‌ها (10^{-18} تا 10^{-5})، پارامترهای بهینه‌ساز به صورت $\beta_1 = 0.9$، $\beta_2 = 0.95$ و $\epsilon = 10^{-15}$ تنظیم شدند [۶۹].

۳. **توقف زودهنگام مبتنی بر تشخیص تکرار (Repetition Early Truncation):** در صورت مشاهده ۳۰۰۰ توکن متوالی با احتمال بالای ۹۹.۰، تولید متوقف می‌شود تا از مصرف بیهوده کانتکست و تزریق گرادیان‌های ناپایدار جلوگیری شود.

۴. **راهبرد توسعه مرحله‌ای پنجره کانتکست (Staged Window Expansion):** طول تولید از ۴۰K آغاز و به ترتیب به ۴۸K، ۵۶K، ۶۴K ۷۲K و نهایتاً ۸۰K ارتقا یافت.

دیاگرام جامع جریان داده، عملیات و انتشار گرادیان

شکل ۶ معماری کامل چرخه داده، ارزیابی، برش ضرایب و پس‌انتشار گرادیان را در الگوریتم **CISPO** نشان می‌دهد.



شکل ۶: دیاگرام جامع جریان داده، محاسبه مزیت درون‌گروهی، برش ضرایب اهمیت و به‌روزرسانی در الگوریتم **CISPO**.

شبه‌کد استاندارد الگوریتم CISPO

الگوریتم ۷، روند اجرایی یادگیری تقویتی مدل با الگوریتم **CISPO** را شرح می‌دهد:

Algorithm 7 Clipped IS-weight Policy Optimization (CISPO)

Input: Policy parameters θ, base reference/sampling model $\pi_{\theta_{\text{old}}}$, group size G, upper IS clip threshold $\epsilon_{\text{high}}^{IS}$, learning rate schedule η, number of off-policy epochs K, repetition early stopping threshold $L_{\text{rep}} = 3000$.
for iteration $= 1, 2, \ldots$ **do**
  Sample a batch of queries $\{q_m\}_{m=1}^{B} \sim \mathcal{D}$.
  Set trajectories buffer $\mathcal{B} \leftarrow \emptyset$.
  for each query q_m **do**
    Sample G independent responses $\{o_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot \mid q_m)$ using rollout engine.
    Apply repetition detection: Halt generation if L_{rep} consecutive tokens have $P(o_t) > 0.99$.
    Compute rewards $R_i = \mathcal{R}(q_m, o_i)$ via rule checkers or Generative Reward Model (GenRM).
    Compute group mean $\mu_R = \frac{1}{G}\sum_{j=1}^{G} R_j$ and std $\sigma_R = \sqrt{\frac{1}{G}\sum_{j=1}^{G}(R_j - \mu_R)^2} + 10^{-8}$.
    Compute advantage for each response: $\hat{A}_i = (R_i - \mu_R)/\sigma_R$.
    Store $(q_m, o_i, \log \pi_{\theta_{\text{old}}}(o_i \mid q_m), \hat{A}_i)$ in buffer $\mathcal{B}$.
  end for
  for epoch $= 1, \ldots, K$ **do**
    Clear gradients: $\nabla_\theta \mathcal{J} \leftarrow 0$.
    Total batch tokens: $N_{\text{total}} = \sum_{(q,o)\in\mathcal{B}} |o|$.
    for each $(q, o_i, \log \pi_{\theta_{\text{old}}}(o_i), \hat{A}_i) \in \mathcal{B}$ **do**
      Compute current token log-probabilities $\log \pi_\theta(o_{i,t} \mid q, o_{i,<t})$ with FP32 LM head.
      Compute token-level importance ratio: $r_{i,t}(\theta) = \exp\left(\log \pi_\theta(o_{i,t}) - \log \pi_{\theta_{\text{old}}}(o_{i,t})\right)$.
      Clip importance weight: $\hat{r}_{i,t}(\theta) = \min\left(r_{i,t}(\theta),\ 1 + \epsilon_{\text{high}}^{IS}\right)$.
      Compute loss component:
        $\mathcal{L}_i(\theta) = -\frac{1}{N_{\text{total}}} \sum_{t=1}^{|o_i|} \text{sg}\left(\hat{r}_{i,t}(\theta)\right) \hat{A}_i \log \pi_\theta(o_{i,t} \mid q, o_{i,<t})$.
      Accumulate gradients: $\nabla_\theta \mathcal{J} \leftarrow \nabla_\theta \mathcal{J} + \nabla_\theta \mathcal{L}_i(\theta)$.
    end for
    Clip global gradient norm: $g \leftarrow \text{clip_norm}(\nabla_\theta \mathcal{J}, \text{max_norm} = 1.0)$.
    Update policy parameters using AdamW ($\beta_1 = 0.9, \beta_2 = 0.95, \epsilon = 10^{-15}$):
      $\theta \leftarrow \theta - \eta \cdot \text{AdamW}(g)$.
  end for
  Synchronize rollout policy: $\theta_{\text{old}} \leftarrow \theta$.
end for

ارزیابی‌های تجربی و تحلیل نتایج بنچمارک‌ها

بر اساس ارزیابی‌های تجربی ارائه‌شده در گزارش فنی **MiniMax-M1** [۵۱]، نتایج به شرح زیر حاصل شده است:

- **شتاب آموزش ۲ برابری نسبت به DAPO:** در آزمون کنترل‌شده روی مدل **Qwen2.5-32B** در بنچمارک ریاضی AIME 2024، الگوریتم **CISPO** با مصرف تنها ٪۵۰ از گام‌های آموزشی به عملکردی برابر با **DAPO** دست یافته و هر دو الگوریتم **GRPO** و **DAPO** را پشت سر گذاشت.
- **کاهش چشمگیر هزینه آموزش:** ترکیب توجه هیبرید با الگوریتم **CISPO** آموزش کامل مدل ۴۵۶ میلیارد پارامتری را روی ۵۱۲ پردازنده H800 در مدت تنها ۳ هفته و با هزینه تقریبی ۵۳۴،۷۰۰ دلار تکمیل نمود.

جدول ۱۷ نتایج جامع مدل‌های آموزش‌دیده با **CISPO** را در مقایسه با پیشرفته‌ترین مدل‌های متن‌باز و تجاری جهان نشان می‌دهد.

جدول ۱۷: ارزیابی مقایسه‌ای مدل‌های **MiniMax-M1** آموزش‌دیده با الگوریتم **CISPO** در بنچمارک‌های استاندارد بین‌المللی [۵۱].

وظیفه / بنچمارک	OpenAI o3	Gemini 2.5 Pro	Claude 4 Opus	DeepSeek-R1	Qwen3-235B	M1-40k	M1-80k
سقف طول تفکر	۱۰۰K	۶۴K	۶۴K	۶۴K	۳۲K	۴۰K	۸۰K
استدلال ریاضی (Mathematics)							
۲۰۲۴ AIME	۶.۹۱	۰.۹۲	۰.۷۶	۴.۹۱	۷.۸۵	۳.۸۳	**۰.۸۶**
۲۰۲۵ AIME	۹.۸۸	۰.۸۸	۵.۷۵	۵.۸۷	۵.۸۱	۶.۷۴	**۹.۷۶**
MATH-۵۰۰	۱.۹۸	۸.۹۸	۲.۹۸	۰.۹۸	۲.۹۶	۰.۹۶	**۸.۹۶**
کدنویسی و مهندسی نرم‌افزار (Coding & Software Engineering)							
LiveCodeBench	۸.۷۵	۱.۷۷	۶.۵۶	۱.۷۳	۹.۶۵	۳.۶۲	**۰.۶۵**
Verified SWE-bench	۱.۶۹	۲.۶۷	۵.۷۲	۶.۵۷	۴.۳۴	۶.۵۵	**۰.۵۶**
بافت فوق طولانی (Long Context Understanding)							
(۱۲۸k) OpenAI-MRCR	۵.۵۶	۸.۷۶	۹.۴۸	۵.۵۱	۷.۲۷	**۱.۷۶**	۴.۷۳
(۱M) OpenAI-MRCR	—	۸.۵۸	—	—	—	**۶.۵۸**	۲.۵۶
LongBench-v۲	۸.۵۸	۰.۶۵	۶.۵۵	۱.۵۲	۱.۵۰	۰.۶۱	**۵.۶۱**
استفاده از ابزار و عامل هوشمند (Agentic Tool Use)							
(Airline) TAU-bench	۰.۵۲	۰.۵۰	۶.۵۹	۵.۵۳	۷.۳۴	۰.۶۰	**۰.۶۲**
(Retail) TAU-bench	۹.۷۳	۰.۶۷	۴.۸۱	۹.۶۳	۶.۵۸	**۸.۶۷**	۵.۶۳

جمع‌بندی

الگوریتم **CISPO** با تشخیص آسیب‌شناسی پنهان در توابع هدف رایج و ارائه راه‌حل مبتنی بر برش وزن اهمیت (IS-Weight Clipping)، موفق شده است ضمن حفظ کامل سیگنال‌های گرادیان در تمام مسیرهای بازتابی و استدلالی، واریانس آموزش برون‌پویا را به طور کامل مهار نماید. این الگوریتم اکنون به عنوان یکی از پایدارترین و اقتصادی‌ترین استانداردهای آموزش یادگیری تقویتی برای ساخت مدل‌های استدلال‌گر بزرگ در مقیاس‌های کانتکست میلیونی شناخته می‌شود.

۷.۷ الگوریتم یادگیری تقویتی با پاداش‌های اعتبارسنجی‌پذیر (Reinforcement Learning with Verifiable Rewards - RLVR)

الگوریتم **RLVR** به عنوان مرحله نهایی خط‌لوله پس‌آموزش در مدل‌های **Tülu 3** توسط لمبرت (Nathan Lambert) و همکاران در موسسه هوش مصنوعی آلن (Ai2) معرفی شد [۷۰]. این الگوریتم با جایگزین ساختن مدل‌های پاداش آماری و تقریبی با توابع اعتبارسنجی قطعی، عملکرد مدل‌های زبانی را در وظایف منطقی، محاسباتی و پیروی دقیق از دستورات به طور چشمگیری ارتقا می‌دهد.

تاریخچه، انگیزه و ضرورت پیدایش الگوریتم

در خطوط لوله متداول هم‌ترازسازی زبانی بر پایه RLHF [۷۱، ۷۲، ۴۰]، بهینه‌سازی سیاست توسط یک مدل پاداش (Reward Model - RM) آموزش‌دیده بر ترجیحات انسانی یا بازخورد هوش مصنوعی (RLAIF) هدایت می‌شود. با این حال، استفاده از مدل پاداش با دو چالش اساسی روبرو است:

- **بهینه‌سازی مفرط و هک پاداش (Reward Hacking / Goodhart's Law):** همان‌طور که گائو و همکاران [۷۳] و سینگهال و همکاران [۷۴] نشان داده‌اند، در گام‌های بالای بهینه‌سازی، خط‌مشی یادگیرنده نقاط کور و سوگیری‌های آماری مدل پاداش (مانند گرایش به پاسخ‌های طولانی یا لحن متملقانه) را به عنوان میان‌بر بهره‌برداری کرده و بدون افزایش کیفیت حقیقی پاسخ، امتیاز پاداش بالایی کسب می‌کند.

- **عدم قطعیت در مسائل با درستی قطعی (Deterministic Verifiability):** در حوزه‌هایی نظیر حل مسائل ریاضیات [۷۵، ۷۶]، اجرای کدهای برنامه‌نویسی و آزمون‌های ساختاریافته دستورپذیری [۷۷]، درستی پاسخ متکی به ترجیحات حسی انسان نبوده بلکه یک گزاره منطقی دودویی (درست یا غلط) است که مدل‌های پاداش استاندارد توان ارزیابی بی‌نقص آن را ندارند.

پیش از این، روش‌هایی مانند **STaR** [۷۸]، **Quiet-STaR** [۷۹]، **TRICE** [۸۰]، بازخورد محیطی در کدنویسی [۸۱] و **VinePPO** [۸۲] تلاش کردند از سیگنال درستی پاسخ استفاده کنند. الگوریتم **RLVR** این دیدگاه را تکامل بخشیده و چارچوبی آنلاین و عمومی بر بستر **PPO** ارائه می‌دهد که بدون نیاز به برچسب‌گذاری تقریب‌گرایانه، مستقیماً پاداش‌های دودویی قطعی را به عنوان سیگنال تقویت زنجیره تفکر مدل زبانی به کار می‌گیرد.

مبانی نظری و اشتقاق‌های کامل ریاضی

فرض کنید یک فرآیند تصمیم‌گیری مارکوف در سطح تولید توکن برای مدل زبانی تعریف شده است. حالت $s_t = (x, y_{<t})$ شامل پرامپت ورودی $x \sim \mathcal{D}$ و توکن‌های پیشین تولیدشده $y_{<t} = (y_1, \ldots, y_{t-1})$ است و عمل $a_t = y_t \in \mathcal{V}$ انتخاب توکن بعدی از واژگان $\mathcal{V}$ است.

تابع هدف بیشینه‌سازی در الگوریتم **RLVR** به صورت زیر تعریف می‌شود:

$$\max_{\pi_\theta} \mathcal{J}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi_\theta(\cdot|x)} \left[R_{\text{RLVR}}(x, y)\right] \tag{۱۱۳}$$

که در آن پاداش کل با تنظیم‌سازی واگرایی کولبک-لایبلر (KL-Divergence) به فرم زیر تعریف می‌گردد:

$$R_{\text{RLVR}}(x, y) = v(x, y) - \beta\, D_{\text{KL}}\left(\pi_\theta(y|x) \parallel \pi_{\text{ref}}(y|x)\right) \tag{۱۱۴}$$

۱. تابع اعتبارسنجی قطعی (Deterministic Verifier Function):
تابع اعتبارسنجی $v(x, y)$ یک مپینگ قطعی روی زوج ورودی-پاسخ است:

$$v(x, y) = \begin{cases} \alpha & \text{اگر استخراج پاسخ با پاسخ صحیح برابر باشد } (\text{Verify}(x, y) = 1) \\ 0 & \text{در غیر این صورت} \end{cases} \tag{۱۱۵}$$

در پیاده‌سازی رسمی [۷۰]، ضریب پاداش $\alpha = 10.0$ تنظیم شده است. همچنین به ازای پاسخ‌هایی که تا انتهای سقف توکن به توکن پایانی EOS نرسند، جریمه $r_{\text{penalty}} = -10.0$ اعمال می‌شود.

۲. تفکیک پاداش در سطح توکن و واگرایی KL:
واگرایی KL توالی کامل $y = (y_1, \ldots, y_T)$ بر حسب ضرب احتمالات شرطی توکن‌ها برابر است با:

$$D_{\text{KL}}(\pi_\theta(y|x) \parallel \pi_{\text{ref}}(y|x)) = \sum_{t=1}^{T} \log\left(\frac{\pi_\theta(y_t \mid x, y_{<t})}{\pi_{\text{ref}}(y_t \mid x, y_{<t})}\right) \tag{۱۱۶}$$

از این رو پاداش اختصاص‌یافته به هر توکن در لحظه t به شکل زیر فرمول‌بندی می‌شود:

$$r_t(x, y) = \begin{cases} -\beta \log\left(\frac{\pi_\theta(y_t|x, y_{<t})}{\pi_{\text{ref}}(y_t|x, y_{<t})}\right) & \text{for} 1 \le t < T \\ v(x, y) - \beta \log\left(\frac{\pi_\theta(y_T|x, y_{<T})}{\pi_{\text{ref}}(y_T|x, y_{<T})}\right) & \text{for} t = T \end{cases} \tag{۱۱۷}$$

۳. محاسبه مزیت با GAE و بهینه‌سازی PPO:
خطای تفاضل زمانی (Temporal Difference Error) برای تابع ارزش V_ϕ با ضریب تخفیف $\gamma = 1.0$ به صورت زیر است:

$$\delta_t^V = r_t + \gamma V_\phi(s_{t+1}) - V_\phi(s_t) \tag{۱۱۸}$$

تابع مزیت تعمیم‌یافته (GAE) با ضریب تعادل $\lambda = 0.95$ محاسبه می‌شود:

$$\hat{A}_t^{\mathsf{GAE}} = \sum_{l=0}^{T-t-1} (\gamma\lambda)^l \delta_{t+l}^V \tag{۱۱۹}$$

مقادیر مزیت در سطح هر دسته به منظور کنترل واریانس نرمال‌سازی می‌شوند:

$$\hat{A}_t^{\mathsf{norm}} = \frac{\hat{A}_t^{\mathsf{GAE}} - \mu(\hat{A}^{\mathsf{GAE}})}{\sigma(\hat{A}^{\mathsf{GAE}}) + 10^{-8}} \tag{۱۲۰}$$

تابع هدف برش‌خورده سیاست (Clipped Surrogate Policy Loss) در **RLVR** به صورت زیر تعریف می‌شود:

$$L^{\mathsf{CLIP}}(\theta) = \hat{\mathbb{E}}_t \left[\min \left(\rho_t(\theta)\hat{A}_t^{\mathsf{norm}},\ \mathsf{clip}(\rho_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t^{\mathsf{norm}} \right) \right] \tag{۱۲۱}$$

که در آن $\rho_t(\theta) = \frac{\pi_\theta(y_t|s_t)}{\pi_{\theta_{\mathsf{old}}}(y_t|s_t)}$ نسبت احتمالات سیاست و $\epsilon = 0.2$ است. تابع هزینه کل شامل خطای پیش‌بینی ارزش $L^{\mathsf{VF}}(\phi) = (V_\phi(s_t) - V_t^{\mathsf{targ}})^2$ به فرم زیر است:

$$\mathcal{L}_{\mathsf{total}}(\theta, \phi) = -L^{\mathsf{CLIP}}(\theta) + c_1 L^{\mathsf{VF}}(\phi) \tag{۱۲۲}$$

که در آن $c_1 = 0.1$ ضریب تلفات تابع ارزش است.

تحلیل‌های تئوری اطلاعات و آمار

سیگنال پاداش $v(x, y)$ در **RLVR** به شدت تنک (Sparse) و ناپیوسته است. بهینه‌سازی مستقیم چنین پاداشی بدون جریمه واگرایی، منجر به فروریزش توزیع احتمالات بر روی الگوهای غیرطبیعی و از دست رفتن روانی زبان می‌گردد. از منظر تئوری اطلاعات، قید واگرایی $D_{\mathsf{KL}}(\pi_\theta \parallel \pi_{\mathsf{ref}})$ سطح آنتروپی زبانی مدل را پایدار نگاه داشته و شعاع جستجوی سیاست را در نزدیکی منیفلد توزیع اولیه حفظ می‌کند.

بررسی‌های تجربی [۷۰] نشان می‌دهد کاهش بیش از حد ضریب β (مانند $\beta = 0.01$) باعث پدیده بهینه‌سازی مفرط و هک قواعد ساختاری می‌شود؛ برای نمونه در آزمون‌های قیود متنی (IFEval)، مدل برای برآورده کردن شرط تکرار کلمه، اقدام به تکرار طوطی‌وار و بی‌معنای همان کلمه می‌کند (شکل ۲۸ در مقاله مرجع). از این رو تنظیم دقیق $\beta \in [0.05, 0.1]$ برای ایجاد تعادل میان درستی منطقی و حفظ گرامر ساختاری الزامی است.

جزئیات زیرساخت محاسباتی و توزیع داده‌ها

۱. مجموعه داده‌های اعتبارسنجی‌پذیر: ترکیب نهایی شامل ۲۹،۹۴۶ پرامپت در سه حوزه اصلی است:

- **مجموعه GSM8K:** شامل ۷،۴۷۳ مسئله ریاضی مقطع ابتدایی با پرامپت هشت‌نمونه‌ای CoT و اعتبارسنجی با استخراج قطعی عدد پایانی.
- **مجموعه MATH:** شامل ۷،۵۰۰ مسئله ریاضی پیشرفته با پرامپت چهارنمونه‌ای و استخراج منعطف به همراه اعتبارسنجی نمادین با بسته SymPy.
- **مجموعه IFEval:** شامل ۱۴،۹۷۳ نمونه ترکیبی همراه با توابع بررسی قاعده دستوری بر پایه عبارات منظم (Regex).

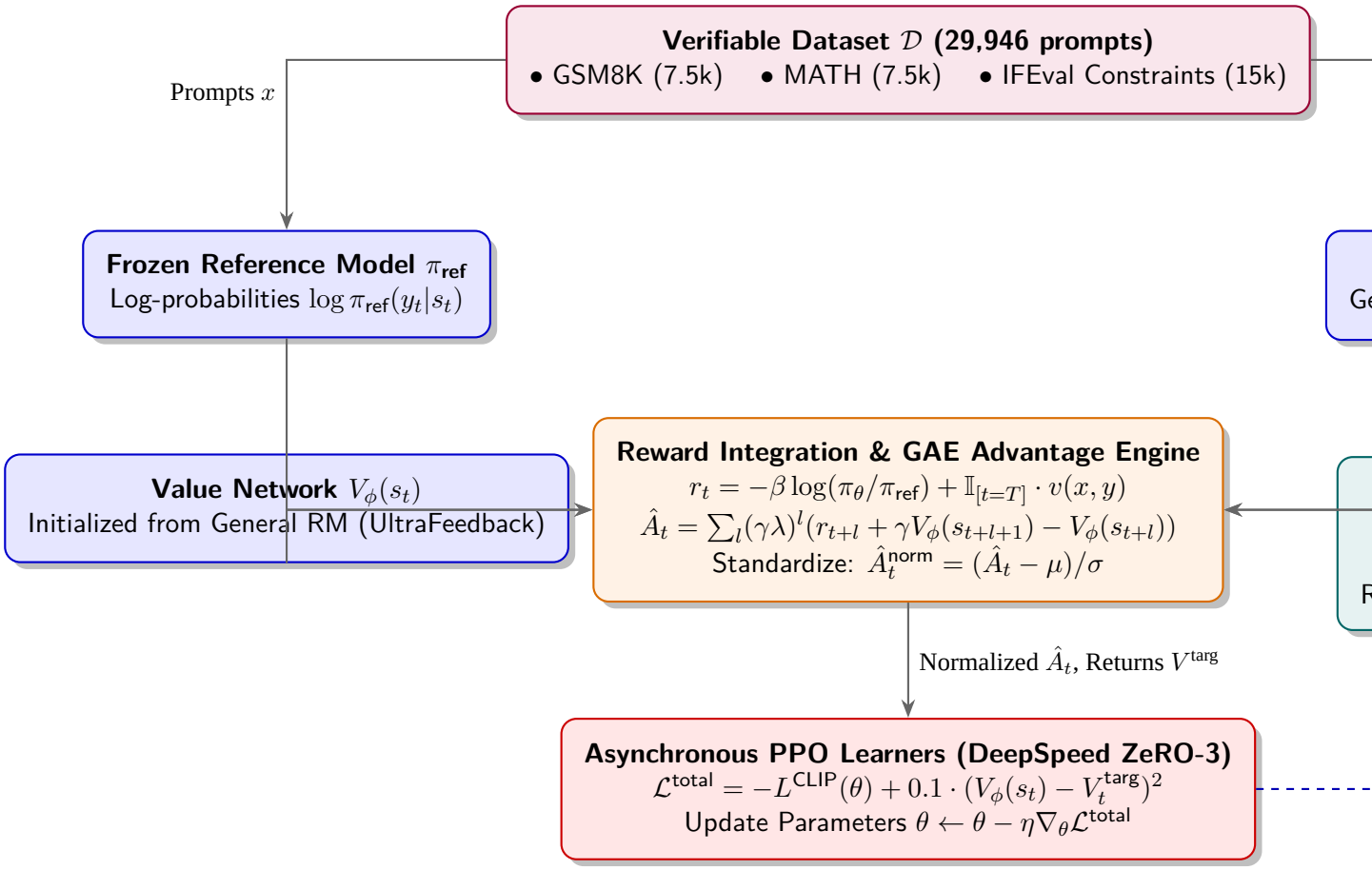
۲. معماری ناهمگام (Asynchronous RL Infrastructure):
برای غلبه بر گلوگاه محاسباتی فاز تولید، موتور استنتاج vLLM [۸۳] با مکانیزم PagedAttention بر روی گره‌های اختصاصی تفکیک شده و مدل‌های یادگیرنده (Learners) با استفاده از فناوری DeepSpeed ZeRO-3 [۸۴] گرادیان‌ها را به طور همزمان محاسبه می‌کنند [۸۵]. وزن‌های جدید پس از هر دور آپدیت با استفاده از برودکست NCCL به موتورهای استنتاج ارسال می‌شوند.

۳. نکات پیاده‌سازی کلیدی:

- **مقداردهی اولیه مدل ارزش:** مدل ارزش V_ϕ مستقیماً از یک مدل پاداش عمومی آموزش‌دیده روی UltraFeedback مقداردهی می‌شود که پایداری همگرایی را نسبت به مقداردهی تصادفی به طور معناداری ارتقا می‌دهد.
- **صفر کردن نرخ دراپ‌اوت (Zero Dropout):** در طول آموزش، Dropout کاملاً غیرفعال است تا محاسبه لاگ-احتمالات در دو فاز نمونه‌برداری و یادگیری کاملاً قطعی و بدون انحراف باقی بماند [۸۶].

دیاگرام جامع جریان محاسباتی الگوریتم RLVR

شکل ۷ جریان داده‌ها، موتور اعتبارسنجی قطعی، نحوه کسر جریمه واگرایی KL و چرخه به‌روزرسانی ناهمگام در الگوریتم **RLVR** را نشان می‌دهد.



شکل ۷: جریان داده‌ها و معماری محاسباتی ناهمگام الگوریتم **RLVR** در چارچوب **Tülu 3**.

شبه‌کد الگوریتم RLVR

شبه‌کد اجرای کامل الگوریتم **RLVR** در الگوریتم ۸ به تصویر کشیده شده است:

Algorithm 8 Reinforcement Learning with Verifiable Rewards (RLVR)

Input: Policy parameters θ_0 (from DPO checkpoint), reference policy π_{ref}, value model parameters ϕ_0 (initialized from RM), verifiable dataset $\mathcal{D}$, reward weight $\alpha = 10.0$, KL penalty β, PPO clip ratio $\epsilon = 0.2$, GAE parameters $\gamma = 1.0, \lambda = 0.95$.
for iteration $= 1, 2, \ldots, N_{\text{steps}}$ **do**
  Sample a batch of prompts with ground truth: $\{(x^{(i)}, y^{(i)}_{\text{gold}})\}_{i=1}^{B} \sim \mathcal{D}$.
  // Step 1: Parallel Rollout Generation via vLLM Engine
  for each prompt $x^{(i)}$ in batch **do**
    Sample completion $y^{(i)} = (y^{(i)}_1, \ldots, y^{(i)}_T) \sim \pi_{\theta_{\text{old}}}(\cdot \mid x^{(i)})$.
    Compute rollout log-probs: $\log \pi_{\theta_{\text{old}}}(y^{(i)}_t \mid x^{(i)}, y^{(i)}_{<t})$.
    Compute reference log-probs: $\log \pi_{\text{ref}}(y^{(i)}_t \mid x^{(i)}, y^{(i)}_{<t})$.
    Compute estimated state values: $V_\phi(s^{(i)}_t)$.
    Evaluate deterministic correctness: $c^{(i)} \leftarrow \text{Verify}(x^{(i)}, y^{(i)}, y^{(i)}_{\text{gold}}) \in \{0, 1\}$.
    Set final verifiable reward: $v(x^{(i)}, y^{(i)}) \leftarrow \alpha \cdot c^{(i)} - 10.0 \cdot \mathbb{I}_{[y^{(i)}_T \neq \text{EOS}]}$.
  end for
  // Step 2: Token-Level Reward Assignment and GAE Computation
  for each token step $t = 1, \ldots, T$ and sequence $i = 1, \ldots, B$ **do**
    Compute token reward:
      $r^{(i)}_t \leftarrow -\beta \left(\log \pi_{\theta_{\text{old}}}(y^{(i)}_t \mid s^{(i)}_t) - \log \pi_{\text{ref}}(y^{(i)}_t \mid s^{(i)}_t) \right) + \mathbb{I}_{[t=T]} \cdot v(x^{(i)}, y^{(i)})$.
    Compute TD residual: $\delta^{V,(i)}_t \leftarrow r^{(i)}_t + \gamma V_\phi(s^{(i)}_{t+1}) - V_\phi(s^{(i)}_t)$.
  end for
  Compute advantage estimates: $\hat{A}^{(i)}_t \leftarrow \sum_{l=0}^{T-t-1} (\gamma\lambda)^l \delta^{V,(i)}_{t+l}$ and target values $V^{\text{targ},(i)}_t \leftarrow \hat{A}^{(i)}_t + V_\phi(s^{(i)}_t)$.
  Standardize advantages across batch: $\hat{A}^{\text{norm},(i)}_t \leftarrow \frac{\hat{A}^{(i)}_t - \mu(\hat{A})}{\sigma(\hat{A}) + 10^{-8}}$.
  // Step 3: Asynchronous Multi-Epoch Learner Optimization
  for epoch $= 1, \ldots, K_{\text{epochs}}$ **do**
    for each minibatch $\mathcal{B}$ **do**
      Compute current log-probabilities: $\log \pi_\theta(y_t \mid s_t)$.
      Compute probability ratio: $\rho_t(\theta) \leftarrow \exp\left(\log \pi_\theta(y_t \mid s_t) - \log \pi_{\theta_{\text{old}}}(y_t \mid s_t)\right)$.
      Compute policy loss:
        $L^{\text{CLIP}}(\theta) \leftarrow \frac{1}{|\mathcal{B}|} \sum_{t \in \mathcal{B}} \min \left(\rho_t(\theta) \hat{A}^{\text{norm}}_t, \text{clip}(\rho_t(\theta), 1-\epsilon, 1+\epsilon) \hat{A}^{\text{norm}}_t \right)$.
      Compute value function loss:
        $L^{\text{VF}}(\phi) \leftarrow \frac{1}{|\mathcal{B}|} \sum_{t \in \mathcal{B}} \left(V_\phi(s_t) - V^{\text{targ}}_t \right)^2$.
      Update parameters via Adam: $\theta, \phi \leftarrow \text{OptimizerStep}\left(\nabla_{\theta,\phi}[-L^{\text{CLIP}}(\theta) + c_1 L^{\text{VF}}(\phi)]\right)$.
    end for
  end for
  Synchronize updated weights: $\theta_{\text{old}} \leftarrow \theta$ via NCCL broadcast to inference nodes.
end for

نتایج تجربی و تحلیل‌های مقایسه‌ای

تأثیر اعمال مرحله **RLVR** پس از بهینه‌سازی **DPO** در ابعاد مختلف مدلی در جدول ۱۸ خلاصه شده است.

جدول ۱۸: ارزیابی مقایسه‌ای مدل‌های آموزش‌دیده با **RLVR** در مقایسه با مبدا **DPO** و خط مبنای **Llama 3.1 Instruct**.[۷۰]

چمارک / معیار ارزیابی	مقیاس ۸B			مقیاس ۷۰B		
	۱.۳ Llama	DPO ۳ Tülu	RLVR ۳ Tülu	۱.۳ Llama	DPO ۳ Tülu	RLVR ۳ Tülu
میانگین کل (Avg.)	۲.۶۲	۴.۶۴	**۸.۶۴**	۴.۷۳	۹.۷۵	**۰.۷۶**
CoT) (۴-shot، **MAT**	۵.۴۲	۰.۴۲	**۷.۴۳**	۴.۵۶	۳.۶۲	**۰.۶۳**
CoT) (۸-shot، **GSM**	۴.۸۳	۳.۸۴	**۶.۸۷**	**۷.۹۳**	۵.۹۳	۵.۹۳
(Strict) **IFEval**	۶.۸۰	۱.۸۱	**۴.۸۲**	**۰.۸۸**	۶.۸۲	۲.۸۳
(LC Winrate) **Alpac**	۲.۲۴	۵.۳۳	**۵.۳۴**	۴.۳۳	۶.۴۹	**۸.۴۹**
(۳-shot) **BigBench-**	۸.۶۲	۸.۶۵	**۰.۶۶**	۸.۷۳	۸.۸۱	**۰.۸۲**
F۱) (۳-shot، **DROP**	۵.۶۱	۵.۶۲	**۶.۶۲**	**۰.۷۷**	۱.۷۴	۳.۷۴

ابرپارامترهای بهینه‌سازی الگوریتم RLVR

جدول ۱۹ مقادیر دقیق ابرپارامترهای بهینه‌شده برای اجرای الگوریتم **RLVR** در مقیاس‌های مختلف را نشان می‌دهد:

جدول ۱۹: پیکربندی ابرپارامترهای آموزش الگوریتم **RLVR** در مدل‌های **Tülu 3**.

ابرپارامتر	مدل ۸B	مدل ۷۰B	مدل ۴۰۵B
نرخ یادگیری سیاست (Policy LR)	3×10^{-7}	1×10^{-7}	1×10^{-7}
ضریب جریمه واگرایی (β)	۰۵.۰	۰۷.۰	۰۵.۰
اندازه دسته موثر (Effective Batch Size)	۲۲۴	۶۴۰	۱۸۵۶
تعداد دور بهینه‌سازی در هر اپیزود (K)	۴	۴	۱
حداکثر طول توکن پاسخ	۲۰۴۸	۲۰۴۸	۱۰۲۴
آستانه برش PPO (ϵ)	۲.۰	۲.۰	۲.۰
ضریب پاداش اعتبارسنجی (α)	۰.۱۰	۰.۱۰	۰.۱۰
جریمه عدم تولید توکن EOS	-۰.۱۰	-۰.۱۰	-۰.۱۰
ضریب تخفیف پاداش (γ)	۰.۱	۰.۱	۰.۱
ضریب پارامتر λ در GAE	۹۵.۰	۹۵.۰	۹۵.۰

جمع‌بندی

الگوریتم **RLVR** با ادغام پاداش‌های قطعی و اثبات‌پذیر در چارچوب گرادیان سیاست مقید به KL، معضل دیرینه هک پاداش در حوزه‌های نیازمند استدلال نمادین و منطقی را برطرف کرده است. این روش بدون فدا کردن توانایی‌های گفتگوی عمومی، مدل‌های خانواده **Tülu 3** را در حل مسائل ریاضی و رعایت دقیق قیود نگارشی به سطحی رقابت‌پذیر با مدل‌های تجاری پیشرو ارتقا داده است.

۸.۷ الگوریتم یادگیری تقویتی با پاداش‌های قابل اعتبارسنجی و بهینه‌سازی گروهی (OlmoRL: RLVR + GRPO)

الگوریتم **OlmoRL** که در جریان توسعه خانواده مدل‌های زبانی باز **Olmo 3** توسط مؤسسه هوش مصنوعی آلن (Ai2) در سال ۲۰۲۵ معرفی شد [۸۷]، یک چارچوب پیشرفته در حوزه پس‌آموزش (Post-Training) استدلالی به شمار می‌رود. این الگوریتم با ارتقا و بازمهندسی الگوریتم بهینه‌سازی خط‌مشی

نسبی گروهی (Group Relative Policy Optimization - GRPO) [۳۹] و تلفیق آن با یادگیری تقویتی مبتنی بر پاداش‌های قابل اعتبارسنجی (Reinforcement Learning with Verifiable Rewards - RLVR)، امکان آموزش پایدار و مقیاس‌پذیر مدل‌های زبانی را برای تولید زنجیره‌های طولانی تفکر (Long Thinking Traces) تا سقف ۳۲۷۶۸ توکن بدون نیاز به شبکه نقاد (Critic Network) فراهم می‌سازد.

تاریخچه، انگیزه و ضرورت پیدایش الگوریتم

توسعه مدل‌های استدلالی متن‌باز در سال‌های اخیر با چالش‌های بنیادین در مرحله یادگیری تقویتی روبرو بوده است:

- **محدودیت‌های ساختاری الگوریتم PPO:** الگوریتم سنتی **PPO** [۳۰] نیازمند نگهداری یک شبکه نقاد مستقل هم‌اندازه با مدل زبانی جهت تخمین تابع ارزش $V(s)$ است. در مدل‌های زبانی بزرگ، این شبکه معادل با دوبرابر شدن مصرف حافظه گرافیکی (VRAM) بوده و تخمین ارزش برای توالی‌های متنی بسیار طولانی با نوسانات شدید و ناپایداری همراه است.
- **نارسایی‌های الگوریتم GRPO اولیه در توالی‌های طولانی:** اگرچه **GRPO** [۳۹] با جایگزینی خط مبنای گروهی (Group Baseline) نیاز به نقاد را مرتفع ساخت، اما در مواجهه با زنجیره‌های تفکر طولانی با معضلاتی نظیر «سوگیری طول» (Length Bias)، «سوگیری دشواری ناشی از نرمال‌سازی انحراف معیار» [۸۸] و «اتلاف محاسبات در دسته‌های با گرادیان صفر» مواجه می‌شود [۵۳].
- **رانش توزیع در استنتاج ناهمگام (Off-Policy Drift):** در سیستم‌های توزیع‌شده مقیاس‌بزرگ، عدم انطباق آماری میان موتور استنتاج سریع نظیر vLLM و موتور آموزش (Learner)، تخمین نسبت اهمیت را مخدوش می‌سازد [۸۹].

الگوریتم **OlmoRL** با هدف برطرف ساختن این موانع، افزایش پایداری در طول دوره‌های طولانی آموزش، بیشینه‌سازی بهره‌وری توان محاسباتی سخت‌افزار (MFU) و گسترش مرزهای استدلال مدل به فراتر از داده‌های نظارت‌شده طراحی شده است.

مبانی نظری و اشتقاق‌های کامل ریاضی

مدل زبانی به عنوان خط‌مشی تصادفی پارامتریزه شده π_θ تعریف می‌شود که به ازای پرامپت ورودی $x = (x_1, \ldots, x_M)$، توالی توکن‌های پاسخ $y = (y_1, \ldots, y_T)$ را به صورت خودبازگشتی تولید می‌کند:

$$\pi_\theta(y|x) = \prod_{t=1}^{T} \pi_\theta(y_t|x, y_{<t}) \tag{۱۲۳}$$

که در آن $y_{<t} = (y_1, \ldots, y_{t-1})$ تاریخچه توکن‌های پیشین است.

هدف یادگیری تقویتی، بیشینه‌سازی امید ریاضی پاداش اسکالر $r(x, y)$ تحت توزیع خط‌مشی است:

$$J(\theta) = \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(\cdot|x)}[r(x, y)] = \sum_x P(x) \sum_y \pi_\theta(y|x) r(x, y) \tag{۱۲۴}$$

با اعمال عملگر گرادیان و استفاده از قاعده مشتق لگاریتمی ($\nabla_\theta \pi = \pi \nabla_\theta \log \pi$):

$$\begin{aligned}
\nabla_\theta J(\theta) &= \sum_x P(x) \sum_y \nabla_\theta \pi_\theta(y|x) r(x, y) \\
&= \sum_x P(x) \sum_y \pi_\theta(y|x) \nabla_\theta \log \pi_\theta(y|x) r(x, y) \\
&= \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(\cdot|x)} \left[\nabla_\theta \log \pi_\theta(y|x) r(x, y)\right]
\end{aligned} \tag{۱۲۵}$$

با باز کردن رابطه برای تک‌تک توکن‌های توالی پاسخ:

$$\nabla_\theta \log \pi_\theta(y|x) = \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(y_t|x, y_{<t}) \tag{۱۲۶}$$

جهت کاهش واریانس، یک خط مبنای مستقل از پاسخ $b(x)$ کسر می‌گردد:

$$\nabla_\theta J(\theta) = \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(\cdot|x)} \left[\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(y_t|x, y_{<t}) \left(r(x,y) - b(x)\right) \right] \tag{۱۲۷}$$

در رویکرد گروهی **GRPO**، به ازای هر پرامپت x، تعداد G پاسخ مستقل $\{y_1, y_2, \ldots, y_G\}$ توسط خط‌مشی پیشین $\pi_{\theta_{\text{old}}}$ تولید می‌شود. الگوریتم **OlmoRL** هفت اصلاح بنیادی را بر این پایه پیاده‌سازی می‌کند:

۱. حذف نرمال‌سازی با انحراف معیار (No Standard Deviation Normalization): در **GRPO** اولیه، مزیت از تقسیم بر انحراف معیار گروه (σ) حاصل می‌شد. طبق تحلیل‌های [۸۸]، این عمل باعث می‌شود در مسائل بسیار سخت یا بسیار آسان که واریانس گروه ناچیز است، مزیت منفجر شده و خط‌مشی دچار بی‌ثباتی گردد. در **OlmoRL** مزیت صرفاً به عنوان اختلاف پاداش با میانگین گروه تعریف می‌شود:

$$A_{i,t} = r(x, y_i) - \text{mean}\left(\{r(x, y_k)\}_{k=1}^{G}\right) = r(x, y_i) - \frac{1}{G}\sum_{k=1}^{G} r(x, y_k) \tag{۱۲۸}$$

۲. نرمال‌سازی در سطح توکن (Token-Level Normalization): به دلیل تنوع طول زنجیره‌های استدلال، میانگین‌گیری در سطح نمونه منجر به وزن‌دهی نابرابر به توکن‌ها می‌شود. **OlmoRL** مجموع خطاهای دسته را بر کل توکن‌های معتبر تمام خروجی‌های گروه تقسیم می‌کند: $\frac{1}{\sum_{i=1}^{G}|y_i|}$ [۵۳].

۳. برش نامتقارن گرادیان (Clip-Higher): نسبت احتمالات گرادیان توکن t به صورت زیر تعریف می‌شود:

$$r_{i,t}(\theta) = \frac{\pi_\theta(y_{i,t}|x, y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t}|x, y_{i,<t})} \tag{۱۲۹}$$

به منظور فراهم‌سازی امکان به‌روزرسانی‌های بزرگ‌تر در جهات مثبت، برش نامتقارن با مقادیر $\varepsilon_{\text{low}} = 0.2$ و $\varepsilon_{\text{high}} = 0.272$ اعمال می‌گردد [۵۳]:

$$\text{clip}\left(r_{i,t}(\theta), 1 - \varepsilon_{\text{low}}, 1 + \varepsilon_{\text{high}}\right) \tag{۱۳۰}$$

۴. تصحیح نمونه‌برداری ترجیحی منقطع (Truncated Importance Sampling - TIS): جهت جبران رانش احتمال میان خروجی موتور استنتاج π_{vllm} و موتور یادگیرنده $\pi(\cdot;\theta_{\text{old}})$، ضریب تصحیح زیر اعمال می‌شود [۸۹]:

$$w_{i,t}^{\text{TIS}} = \min\left(\frac{\pi(y_{i,t}|x, y_{i,<t}; \theta_{\text{old}})}{\pi_{\text{vllm}}(y_{i,t}|x, y_{i,<t}; \theta_{\text{old}})},\ \rho\right) \tag{۱۳۱}$$

که در آن ρ سقف مجاز بازوزن‌دهی است.

۵. حذف جریمه صریح واگرایی کولبک-لایبلر (No-KL Objective): ترم جریمه صریح D_{KL} از تابع هزینه حذف شده تا آزادی اکتشاف مدل در یافتن زنجیره‌های استدلال جدید محدود نگردد [۵۳، ۸۸].

۶. فیلتر کردن سیگنال‌های گرادیان صفر (Zero Gradient Signal Filtering): گروه‌هایی که پاداش تمام نمونه‌های آن‌ها یکسان باشد ($A_{i,t} = 0$)، پیش از گام پس‌انتشار به طور کامل حذف می‌گردند.

۷. نمونه‌برداری فعال پویا (Active Dynamic Sampling): صف توزیع‌شده به صورت بلادرنگ پرامپت‌های فیلترشده را جایگزین می‌کند تا اندازه دسته آموزشی همواره ثابت بماند.

تابع هدف نهایی الگوریتم OlmoRL

با تجمیع روابط فوق، تابع هدف نهایی بهینه‌سازی در **OlmoRL** به صورت رابطه ریاضی زیر تعریف می‌شود:

(۱۳۲)

$$J(\theta) = \frac{1}{\sum_{i=1}^{G}|y_i|}\sum_{i=1}^{G}\sum_{t=1}^{|y_i|}\min\left(\frac{\pi(y_{i,t}|x, y_{i,<t};\theta_{\text{old}})}{\pi_{\text{vllm}}(y_{i,t}|x, y_{i,<t};\theta_{\text{old}})},\ \rho\right)\cdot\min\left(r_{i,t}(\theta)A_{i,t},\ \text{clip}(r_{i,t}(\theta), 1-\varepsilon_{\text{low}}, 1+\varepsilon_{\text{high}})A_{i,t}\right)$$

معماری سیستم پاداش‌های قابل اعتبارسنجی (RLVR) و اعتبارسنج‌ها

الگوریتم **OlmoRL** ساختار اعتبارسنجی خودکار را به چهار دامنه مختلف تعمیم می‌دهد:

- **دامنه ریاضیات (Math Verifier):** پاسخ مدل از طریق تطبیق عبارات و ارزیابی هم‌ارزی نمادین جبری توسط کتابخانه SymPy اعتبارسنجی می‌شود ($r \in \{0, 1\}$).
- **دامنه کدنویسی (Code Verifier):** کد تولیدی در محیط‌های ایزوله و موازی بر بستر AWS Lambda [۹۰] اجرا شده و پاداش بر مبنای نرخ قبولی تست‌ها (Pass Rate) یا قبولی کامل تعیین می‌گردد.
- **پیروی دقیق از دستورالعمل (Precise Instruction Following):** ارزیابی پایبندی به قیود ساختاری (مانند تعداد کلمات، پاراگراف‌ها و فرمت خروجی) از طریق توابع قطعی چارچوب IF-RLVR [۹۱] انجام می‌پذیرد.
- **مکالمه عمومی (General Chat):** ارزیابی کیفی پاسخ‌های متنی آزاد با استفاده از مدل داور **Qwen3-32B** در بازه مقداری $[0, 1]$ صورت می‌گیرد.

تحلیل‌های آماری، تئوری اطلاعات و پویایی‌های آموزش

- **پویایی طول پاسخ‌ها (Length Dynamics):** در مراحل اولیه آموزش، طول توالی تفکر مدل کاهش اندکی می‌یابد و سپس با کشف مسیرهای استدلالی پیچیده نظیر خودآزمایی و بازگشت از خطا، طول پاسخ‌ها به تدریج افزایش یافته و به سقف ۳۲K توکن نزدیک می‌شود.
- **تنظیم‌کنندگی ضمنی آموزش چنددامنه‌ای (Multi-Domain Regularization):** آموزش همزمان بر روی ترکیب داده‌های ریاضی، کد، دستورالعمل و چت، پاداش ظاهری فاز آموزش را در مقایسه با آموزش تک‌دامنه‌ای اندکی کاهش می‌دهد، اما مانع از بیش‌برازش و هک پاداش (Reward Hacking) شده و عملکرد تعمیم‌یافته مدل در بنچمارک‌های نهایی را تثبیت می‌کند.
- **اثبات عدم آلودگی داده‌ها با آزمون کنترل منفی (Spurious Rewards):** در یک آزمایش کنترل منفی بر پایه پروتکل [۹۲]، پاداش‌های تصادفی به مدل تخصیص یافت؛ عدم مشاهده هیچ‌گونه بهبود در بنچمارک‌ها اثبات کرد که پیشرفت‌های **OlmoRL** ناشی از یادگیری واقعی استدلال است و نه فعال‌سازی حافظه آلوده.

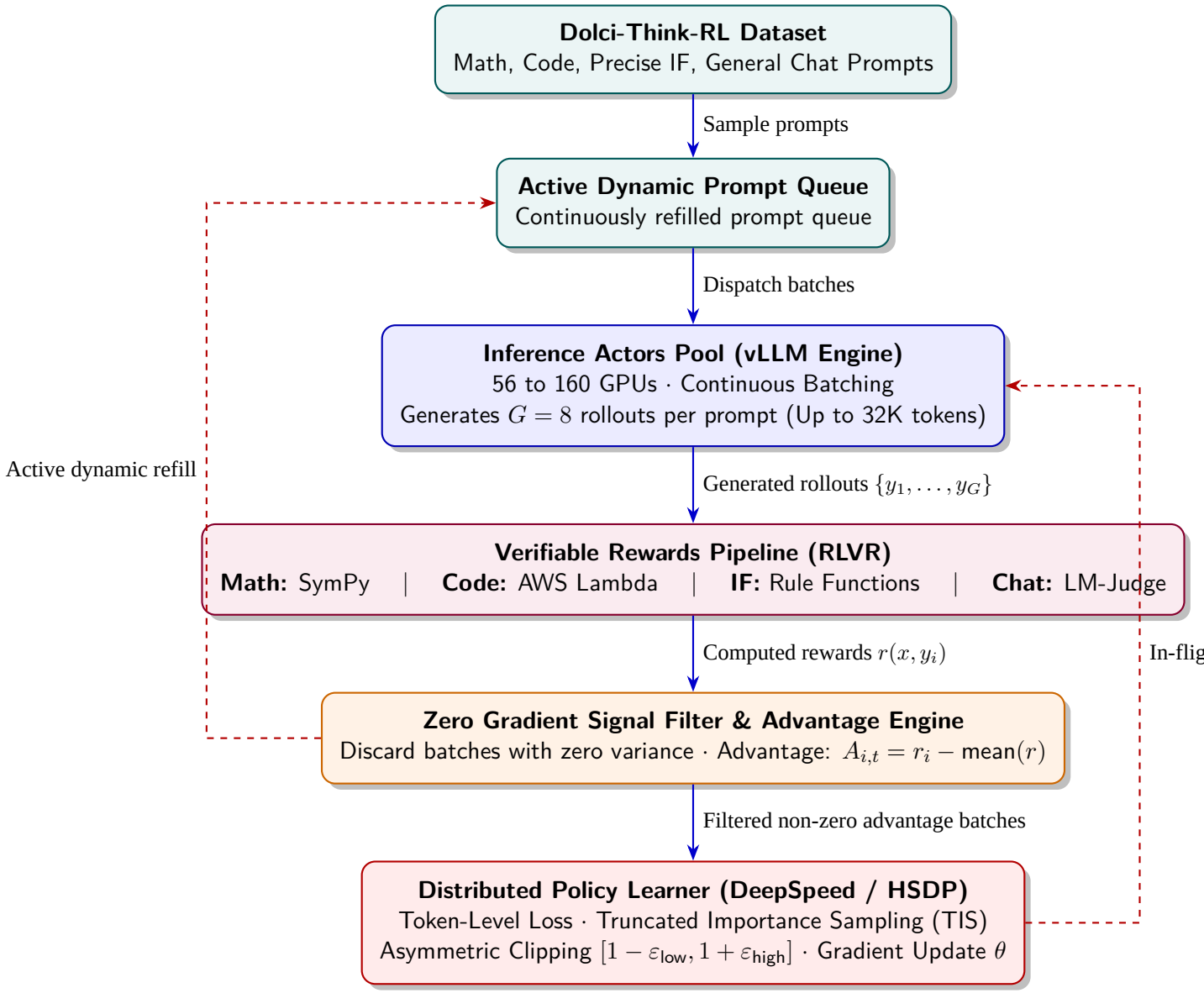
معماری زیرساخت محاسباتی و مهندسی سیستم

اجرای **OlmoRL** برای زنجیره‌های استدلال تا ۳۲K توکن نیازمند نوآوری‌های زیرساختی متعددی بوده است:

- **دسته‌بندی پیوسته (Continuous Batching):** با توجه به میانگین طول توالی ۱۴۶۲۸ توکن و سقف ۳۲۷۶۸ توکن، استفاده از دسته‌بندی ایستا منجر به اتلاف محاسباتی معادل $\frac{L_{\text{max}}-L_{\text{avg}}}{L_{\text{max}}} \approx 54\%$ می‌شد. دسته‌بندی پیوسته این اتلاف را به صفر رسانده است.
- **به‌روزرسانی‌های پروازی وزن‌ها (In-flight Weight Updates):** با الهام از PipelineRL [۹۳]، وزن‌های جدید مدل یادگیرنده بدون متوقف ساختن موتور استنتاج و بدون تخلیه حافظه کَش KV-Cache به بازیگران تزریق می‌شود که شتاب ۴ برابری در توان عملیاتی ایجاد کرده است.

دیاگرام معماری و جریان داده‌های الگوریتم

شکل ۸ ساختار تعاملی میان استخر بازیگران، صف پویا، اعتبارسنج‌های خودکار، فیلتر داده‌ها و یادگیرنده توزیع‌شده در الگوریتم **OlmoRL** را نمایش می‌دهد.



شکل ۸: معماری جامع زیرساخت توزیع‌شده، پایپ‌لاین اعتبارسنجی پاداش‌ها و جریان بهینه‌سازی در الگوریتم **OlmoRL**.

شبه‌کد الگوریتم OlmoRL

شبه‌کد فرآیند کامل نمونه‌برداری پویا، اعتبارسنجی، فیلتر داده‌ها و به‌روزرسانی پارامترها در الگوریتم ۹ ارائه شده است:

Algorithm 9 OlmoRL: Group Relative Policy Optimization with Verifiable Rewards

Input: Initial policy θ_0, inference engine π_{vllm}, group size $G = 8$, target batch size B, clipping bounds $(\varepsilon_{\text{low}}, \varepsilon_{\text{high}})$, TIS cap ρ, learning rate η.
for step $= 1, 2, \ldots$ **do**
  Initialize batch container $\mathcal{B} \leftarrow \emptyset$.
  while $|\mathcal{B}| < B$ **do**
    Fetch prompt x from active queue.
    Sample G completions $\{y_1, \ldots, y_G\} \sim \pi_{\text{vllm}}(\cdot|x)$ with length up to 32K.
    for each completion y_i **do**
      Compute reward $r(x, y_i)$ via domain-specific verifier (SymPy, Sandbox, Rules, or Judge).
    end for
    Compute group mean reward: $\bar{r} = \frac{1}{G}\sum_{k=1}^{G} r(x, y_k)$.
    if $r(x, y_1) == r(x, y_2) == \cdots == r(x, y_G)$ **then**
      continue ▷ Zero Gradient Signal Filtering
    end if
    Compute advantages: $A_i = r(x, y_i) - \bar{r}, \quad \forall i \in \{1, \ldots, G\}$.
    Add tuple $(x, \{y_i\}, \{A_i\})$ to $\mathcal{B}$.
  end while
  Compute total valid tokens: $N_{\text{tokens}} = \sum_{(x,\{y_i\},\cdot)\in\mathcal{B}} \sum_{i=1}^{G} |y_i|$.
  Zero model gradients: $\nabla_\theta J \leftarrow 0$.
  for each $(x, \{y_i\}, \{A_i\}) \in \mathcal{B}$ **do**
    for $i = 1$ **to** G **do**
      for $t = 1$ **to** $|y_i|$ **do**
        Compute ratio: $r_{i,t}(\theta) = \frac{\pi_\theta(y_{i,t}|x, y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t}|x, y_{i,<t})}$.
        Compute TIS weight: $w_{i,t}^{\text{TIS}} = \min\left(\frac{\pi(y_{i,t}|x, y_{i,<t}; \theta_{\text{old}})}{\pi_{\text{vllm}}(y_{i,t}|x, y_{i,<t}; \theta_{\text{old}})}, \rho\right)$.
        Compute clipped surrogate:
          $\text{obj}_{i,t} = \min\left(r_{i,t}(\theta)A_i,\ \text{clip}(r_{i,t}(\theta), 1 - \varepsilon_{\text{low}}, 1 + \varepsilon_{\text{high}})A_i\right)$.
        Accumulate loss: $\mathcal{L} \leftarrow \mathcal{L} - \frac{1}{N_{\text{tokens}}} w_{i,t}^{\text{TIS}} \cdot \text{obj}_{i,t}$.
      end for
    end for
  end for
  Perform gradient step: $\theta \leftarrow \theta - \eta\, \text{Adam}(\nabla_\theta \mathcal{L})$.
  Asynchronously update inference actor weights: $\pi_{\text{vllm}} \leftarrow \theta$ without clearing KV cache.
end for

نتایج تجربی، ارزیابی بنچمارک‌ها و تحلیل ابلیشن

جدول ۲۰ تاثیر مولفه‌های مهندسی **OlmoRL** بر شتاب محاسباتی و جدول ۲۱ عملکرد مدل پرچم‌دار **Olmo 3 Think-32B** را در مقایسه با برترین مدل‌های هم‌رده نمایش می‌دهند.

ابرپارامترهای استاندارد گزارش‌شده در مقاله

جدول ۲۲ مقادیر دقیق ابرپارامترهای آموزشی الگوریتم **OlmoRL** را برای مدل‌های مختلف نشان می‌دهد:

جدول ۲۰: تحلیل ابلیشن زیرساخت **OlmoRL** و تاثیر آن بر سرعت آموزش و بهره‌وری سخت‌افزار [۸۷].

پیکربندی زیرساخت	مجموع توکن‌ها (Mtok)	سرعت (Tokens/sec)	MFU (%)	MBU (%)
زیرساخت پایه **OLMo 2**	۳۴.۶	۸۸۱	۳۰.۰	۹۰.۱۲
بندی پیوسته (Continuous Batching)	۰۲.۷	۹۷۵	۳۳.۰	۲۹.۱۴
ی نخ‌های پردازشی (Better Threading)	۷۷.۹	۱۳۵۸	۴۶.۰	۸۹.۱۹
به‌روزرسانی پروازی وزن‌ها (OlmoRL)	**۲۳.۲۱**	**۲۹۴۹**	**۰۱.۱**	**۲۱.۴۳**

جدول ۲۱: مقایسه عملکرد مدل استدلالی **Olmo 3 Think-32B** در بنچمارک‌های استاندارد [۸۷].

بنچمارک	Olmo 3 SFT	Olmo 3 DPO	Olmo 3 Think	Qwen 3 32B	Qwen 2.5 32B	Gemma 3 27B
MATH	۶.۹۵	۹.۹۵	**۱.۹۶**	۴.۹۵	۲.۸۰	۴.۸۷
۲۰۲۴ AIME	۵.۷۳	۰.۷۶	**۸.۷۶**	۸.۸۰	۷.۱۵	۹.۲۸
۲۰۲۵ AIME	۲.۶۶	۷.۷۰	**۵.۷۲**	۹.۷۰	۴.۱۳	۹.۲۲
v۳ LiveCod	۸.۷۵	۹.۸۱	**۵.۸۳**	۲.۹۰	۹.۴۹	۰.۳۹
HumanEva	۰.۹۰	۶.۹۱	**۴.۹۱**	۲.۹۱	۶.۸۲	۲.۷۹
BigBenchH	۸.۸۸	۱.۸۹	**۸.۸۹**	۶.۹۰	۹.۸۰	۴.۸۲
IFEval	۹.۸۳	۶.۸۰	**۰.۸۹**	۵.۸۶	۹.۸۱	۴.۸۵

جدول ۲۲: ابرپارامترهای نهایی آموزش یادگیری تقویتی **OlmoRL** در مدل‌های خانواده **Olmo 3** [۸۷].

ابرپارامتر	Think 7B	Think 32B	Instruct 7B	RL-Zero 7B
تعداد کل پرامپت‌ها	۱۰۴،۸۶۹	۱۰۴،۸۶۹	۱۷۱،۹۵۰	۱۳،۳۱۴
نرخ یادگیری (Learning Rate)	1.0×10^{-6}	2.0×10^{-6}	1.0×10^{-6}	1.0×10^{-6}
برنامه نرخ یادگیری	Constant	Constant	Constant	Constant
اندازه گروه (G)	۸	۸	۸	۸
حداکثر طول پرامپت	۲،۰۴۸	۲،۰۴۸	۲،۰۴۸	۲،۰۴۸
حداکثر طول پاسخ	۳۲،۷۶۸	۳۲،۷۶۸	۸،۱۹۲	۱۶،۳۸۴
آستانه برش نامتقارن ($\varepsilon_{\text{low}}, \varepsilon_{\text{high}}$)	(۲.۰، ۲۷۲.۰)	(۲.۰، ۲۷۲.۰)	(۲.۰، ۲۷۲.۰)	(۲.۰، ۲۷۲.۰)
سقف بازوزن‌دهی ρ در TIS	–	۰.۲	۲.۰	۰.۲
دمای نمونه‌برداری	۰.۱	۰.۱	۰.۱	۰.۱
تعداد کارت‌های یادگیرنده / بازیگر	۱۶ / ۵۶	۶۴ / ۱۶۰	۸ / ۵۶	۸ / ۶۴

جمع‌بندی

الگوریتم **OlmoRL** با بازنگری فرمولاسیون ریاضی محاسبه مزیت، حذف انحراف معیار جهت رفع سوگیری دشواری، تلفیق اعتبارسنجی چنددامنه‌ای قطعی و نوآوری در همگام‌سازی پروازی وزن‌ها، زیرساختی پایدار و فوق‌العاده سریع برای استخراج رفتارهای استدلالی پدیدار شونده در مدل‌های زبانی باز پدید آورده است. این چارچوب توانسته است مدل **Olmo 3 Think-32B** را به قدرتمندترین مدل استدلالی کاملاً متن‌باز تبدیل کند.

۸ منابع

[1] Houyi Li, Wenzhen Zheng, Qiufeng Wang, Zhenyu Ding, Haoying Wang, Zili Wang, Shijie Xuyang, Ning Ding, Shuigeng Zhou, Xiangyu Zhang, et al. Predictable scale: Part ii, farseer–a refined scaling law in large language models. *arXiv preprint arXiv:2506.10972*, 2025.

[2] Zhengyu Chen, Siqi Wang, Teng Xiao, Yudong Wang, Shiqi Chen, Xunliang Cai, Junxian He, and Jingang Wang. Revisiting scaling laws for language models: The role of data quality and training strategies. *arXiv preprint*, 2025.

[3] Sean McLeish, John Kirchenbauer, David Yu Miller, Siddharth Singh, Abhinav Bhatele, Micah Goldblum, Ashwinee Panda, and Tom Goldstein. Gemstones: A model suite for multi-faceted scaling laws. *arXiv preprint arXiv:2502.06857*, 2025.

[4] Song Bian, Tao Yu, Shivaram Venkataraman, and Youngsuk Park. Scaling laws meet model architecture: Toward inference-efficient llms. *arXiv preprint arXiv:2510.18245*, 2025.

[5] Yizhou Liu, Sara Kangaslahti, Ziming Liu, and Jeff Gore. Inverse depth scaling from most layers being similar. *arXiv preprint arXiv:2602.05970*, 2026.

[6] Wenjie Sun, Jinning Yang, Shuai Zhang, and Mengnan Du. Law of neural interaction: Depth–width shape, interaction efficiency, and generalization. *arXiv preprint arXiv:2605.27989*, 2026.

[7] Nandan Kumar Jha and Brandon Reagen. Spectral scaling laws in language models: How effectively do feed-forward networks use their latent space? *arXiv preprint arXiv:2510.00537*, 2025.

[8] Maximilian Beck, Kajetan Schweighofer, Sebastian Böck, Sebastian Lehner, and Sepp Hochreiter. xlstm scaling laws: Competitive performance with linear time-complexity. *arXiv preprint arXiv:2510.02228*, 2026.

[9] Anonymous. On the recall scaling laws in mamba: A theoretical and mechanistic study via hashing. *arXiv preprint*, 2026.

[10] Changxin Tian, Kunlong Chen, Jia Liu, Ziqi Liu, Zhiqiang Zhang, and Jun Zhou. Towards greater leverage: Scaling laws for efficient mixture-of-experts language models. *arXiv preprint arXiv:2507.17702*, 2025.

[11] Samira Abnar, Harshay Shah, Dan Busbridge, Alaaeldin El-Nouby, Josh Susskind, and Vimal Thilak. Parameters vs flops: Scaling laws for optimal sparsity for mixture-of-experts language models. *arXiv preprint arXiv:2501.12370*, 2025.

[12] Jan Ludziejewski, Maciej Pióro, Jakub Krajewski, Maciej Stefaniak, Michał Krutul, Jan Małaśnicki, Marek Cygan, Piotr Sankowski, Kamil Adamczewski, Piotr Miłoś, et al. Joint moe scaling laws: Mixture of experts can be memory efficient. *arXiv preprint arXiv:2502.05172*, 2025.

[13] Seng Pei Liew, Takuya Kato, and Sho Takase. Scaling laws for upcycling mixture-of-experts language models. *arXiv preprint arXiv:2502.03009*, 2025.

[14] Anirudh Subramanyam, Yuxin Chen, and Robert L Grossman. Scaling laws revisited: Modeling the role of data quality in language model pretraining. *arXiv preprint arXiv:2510.03313*, 2026.

[15] Mustafa Shukor, Louis Bethune, Dan Busbridge, David Grangier, Enrico Fini, Alaaeldin El-Nouby, and Pierre Ablin. Scaling laws for optimal data mixtures. *arXiv preprint arXiv:2507.09404*, 2025.

[16] Francesco Cagnetta, Alessandro Favero, Antonio Sclocchi, and Matthieu Wyart. Scaling laws and representation learning in simple hierarchical languages: Transformers vs. convolutional architectures. *arXiv preprint arXiv:2505.07070*, 2025.

[17] William Held, David Hall, Percy Liang, and Diyi Yang. Relative scaling laws for llms. *arXiv preprint arXiv:2510.24626*, 2026.

[18] Baoqing Yue, Jinyuan Zhou, Zixi Wei, Jingtao Zhan, Qingyao Ai, and Yiqun Liu. Relative-based scaling law for neural language models. *arXiv preprint arXiv:2510.20387*, 2025.

[19] Chaojun Xiao, Jie Cai, Weilin Zhao, Biyuan Lin, Guoyang Zeng, Jie Zhou, Zhi Zheng, Xu Han, Zhiyuan Liu, and Maosong Sun. Densing law of llms. *Nature Machine Intelligence*, pages 1–11, 2025.

[20] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2nd edition, 2018.

[21] Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. *International Conference on Machine Learning (ICML)*, 2:267–274, 2002.

[22] Richard S. Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems (NIPS)*, volume 12, pages 1057–1063, 2000.

[23] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3-4):229–256, 1992.

[24] J. Andrew Bagnell and Jeff Schneider. Covariant policy search. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1019–1024, 2003.

[25] Sham Kakade. A natural policy gradient. In *Advances in Neural Information Processing Systems (NIPS)*, volume 14, pages 1531–1538, 2002.

[26] John Schulman, Sergey Levine, Philipp Moritz, Michael Jordan, and Pieter Abbeel. Trust region policy optimization. *International Conference on Machine Learning (ICML)*, pages 1889–1897, 2015.

[27] Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized markov decision processes. In *International Conference on Machine Learning (ICML)*, volume 97 of *PMLR*, pages 2160–2169, 2019.

[28] Bruno Scherrer. Approximate policy iteration schemes: A comparison. In *International Conference on Machine Learning (ICML)*, volume 32 of *JMLR W&CP*, pages 1314–1322, 2014.

[29] Bruno Scherrer and Boris Lesner. On the use of non-stationary policies for stationary infinite-horizon markov decision processes. In *Advances in Neural Information Processing Systems (NIPS)*, volume 25, 2012.

[30] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[31] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[32] Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*, 2015.

[33] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

[34] Yan Duan, Xi Chen, Rein Houthooft, John Schulman, and Pieter Abbeel. Benchmarking deep reinforcement learning for continuous control. *International Conference on Machine Learning (ICML)*, pages 1329–1338, 2016.

[35] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. *International Conference on Machine Learning (ICML)*, pages 1928–1937, 2016.

[36] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[37] John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015.

[38] Nicolas Heess, Srinivasan Sriram, Jay Lemmon, Josh Merel, Greg Wayne, Yuval Tassa, Tom Erez, Ziyu Wang, SM Eslami, Martin Riedmiller, et al. Emergence of locomotion behaviours in rich environments. *arXiv preprint arXiv:1707.02286*, 2017.

[39] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[40] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[41] John Schulman. Approximating kl divergence. *Online blog: http://joschu.net/blog/kl-approx.html*, 2020.

[42] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

[43] Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deyu Chen, Y Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. *arXiv preprint arXiv:2312.08935*, 2023.

[44] Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Chuanqi Tan, and Chang Zhou. Scaling relationship on learning mathematical reasoning with large language models. *arXiv preprint arXiv:2308.01825*, 2023.

[45] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2023.

[46] Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

[47] Team Qwen. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[48] Team Qwen. Qwq-32b: Embracing the power of reinforcement learning. https://qwenlm.github.io/blog/qwq-32b/, 2025.

[49] OpenAI. Learning to reason with llms. https://openai.com/index/learning-to-reason-with-llms/, 2024.

[50] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[51] MiniMax. Minimax-m1: Scaling test-time compute efficiently with lightning attention. *arXiv preprint arXiv:2506.13585*, 2025.

[52] Chujie Zheng, Pei Ke, Zheng Zhang, and Minlie Huang. Click: Controllable text generation with sequence likelihood contrastive learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 65–78, 2023.

[53] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

[54] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[55] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[56] Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv:2409.19256*, 2024.

[57] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

[58] Chang Gao, Chujie Zheng, Xiong-Hui Chen, Kai Dang, Shixuan Liu, Bowen Yu, An Yang, Shuai Bai, Jingren Zhou, and Junyang Lin. Soft adaptive policy optimization. *arXiv preprint arXiv:2511.20347*, 2025.

[59] Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

[60] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[61] Team Qwen. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[62] Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*, 2024.

[63] Bill Yuchen Lin, Ronan Le Bras, Kyle Richardson, Ashish Sabharwal, Radha Poovendran, Peter Clark, and Yejin Choi. Zebralogic: On the scaling limits of llms for logical reasoning. *arXiv preprint arXiv:2502.01100*, 2025.

[64] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 95095–95169, 2024.

[65] Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. Swe-bench: Can language models resolve real-world github issues? In *International Conference on Learning Representations (ICLR)*, 2024.

[66] Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, et al. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*, 2025.

[67] Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, et al. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.01939*, 2025.

[68] Zhen Qin, Weigao Sun, Dong Li, Xuyang Shen, Weixuan Sun, and Yiran Zhong. Lightning attention-2: A free lunch for handling unlimited sequence lengths in large language models. *arXiv preprint arXiv:2401.04658*, 2024.

[69] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *International Conference on Learning Representations (ICLR)*, 2019.

[70] Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Xinxi Lyu, et al. Tülu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

[71] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30, 2017.

[72] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

[73] Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. *International Conference on Machine Learning*, pages 10835–10866, 2023.

[74] Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. A long way to go: Investigating length correlations in rlhf. *arXiv preprint arXiv:2403.07862*, 2024.

[75] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[76] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

[77] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

[78] Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, 2022.

[79] Eric Zelikman, Georges Harik, Yining Shao, Varuna Jayasiri, Nick Haber, and Noah D Goodman. Quiet-star: Language models can teach themselves to think before speaking. *arXiv preprint arXiv:2403.09629*, 2024.

[80] Matthew D Hoffman, Du Phan, David Dohan, Sholto Douglas, Tuan Anh Le, Aaron Parisi, Pavel Sountsov, Charles Sutton, Sharad Vikram, and Rif A Saurous. Training chain-of-thought via latent-variable inference. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[81] Jonas Gehring, Kunhao Zheng, Jade Copet, Valerio Mella, Taco Cohen, and Gabriel Synnaeve. Rlef: Grounding code llms in execution feedback with reinforcement learning. *arXiv preprint arXiv:2410.02089*, 2024.

[82] Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron Courville, and Nicolas Le Roux. Vineppo: Unlocking rl potential for llm reasoning through refined credit assignment. *arXiv preprint arXiv:2410.01679*, 2024.

[83] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th ACM SIGOPS Symposium on Operating Systems Principles*, pages 611–626, 2023.

[84] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE, 2020.

[85] Michael Noukhovitch, Shengyi Huang, Sophie Xhonneux, Arian Hosseini, Rishabh Agarwal, and Aaron Courville. Asynchronous rlhf: Faster and more efficient off-policy rl for language models. *arXiv preprint arXiv:2410.18252*, 2024.

[86] Shengyi Huang, Michael Noukhovitch, Arian Hosseini, Kashif Rasul, Weixun Wang, and Lewis Tunstall. The n+ implementation details of rlhf with ppo: A case study on tl;dr summarization. In *First Conference on Language Modeling*, 2024.

[87] Olmo Team, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. Olmo 3: Fully open language and thinking models. *arXiv preprint*, 2025.

[88] Z. Liu, C. Chen, W. Li, P. Qi, T. Pang, C. Du, W. S. Lee, and M. Lin. Understanding r1-zero-like training: A critical perspective. *Conference on Language Modeling (COLM)*, 2025.

[89] F. Yao, L. Liu, D. Zhang, C. Dong, J. Shang, and J. Gao. Your efficient rl framework secretly brings you off-policy rl training. *Notion Technical Report*, 2025.

[90] H. Zeng, J. Yang, Y. Zhang, B. Yu, S. Wang, Z. Liu, M. Sun, and T. Liu. Acecoder: Acing coder rl via automated test-case synthesis. *arXiv preprint arXiv:2502.01718*, 2025.

[91] Valentina Pyatkin, Saumya Malik, Victoria Graf, Hamish Ivison, Shengyi Huang, Pradeep Dasigi, Nathan Lambert, and Hannaneh Hajishirzi. Generalizing verifiable instruction following. *arXiv preprint arXiv:2507.02833*, 2025.

[92] R. Shao, S. S. Li, R. Xin, S. Geng, Y. Wang, S. Oh, S. S. Du, N. Lambert, S. Min, R. Krishna, et al. Spurious rewards: Rethinking training signals in rlvr. *arXiv preprint arXiv:2506.10947*, 2025.

[93] A. Piché, E. Kamaloo, R. Pardinas, and D. Bahdanau. Pipelinerl: Faster on-policy reinforcement learning for long sequence generation. *arXiv preprint arXiv:2509.19128*, 2025.

[94] Ziyu Wang, Victor Bapst, Nicolas Heess, Volodymyr Mnih, Remi Munos, Koray Kavukcuoglu, and Nando de Freitas. Sample efficient actor-critic with experience replay. *arXiv preprint arXiv:1611.01224*, 2016.

[95] Xing Chen, Dongcui Diao, Hechang Chen, Hengshuai Yao, Haiyin Piao, Zhixiao Sun, Zhiwei Yang, Randy Goebel, Bei Jiang, and Yi Chang. The sufficiency of off-policyness and soft clipping: Ppo is still insufficient according to an off-policy measure. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 7078–7086, 2023.

[96] Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik R Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. In *International Conference on Learning Representations (ICLR)*, 2025.

[97] Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, et al. Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2412.15204*, 2024.

[98] Junteng Liu, Yuanxiang Fan, Zhuo Jiang, Han Ding, Yongyi Hu, Chi Zhang, et al. Synlogic: Synthesizing verifiable reasoning data at scale for learning logical reasoning and beyond. *arXiv preprint arXiv:2505.19641*, 2025.

[99] J. Andrew Bagnell, Sham M. Kakade, Andrew Y. Ng, and Jeff Schneider. Policy search by dynamic programming. In *Advances in Neural Information Processing Systems (NIPS)*, volume 16, 2003.

[100] Bruno Scherrer and Matthieu Geist. Local policy search in a convex space and conservative policy iteration as boosted policy search. In *European Conference on Machine Learning (ECML)*, volume 8725 of *LNAI*, pages 35–50, 2014.