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) ۳۰
۲.۱ روش (Method) ۴۶
۱.۲.۱ پیش‌آموزش (Pre-Training) ۴۶
۲.۲.۱ پس‌آموزش (Post-Training) ۴۶

۲ ساختار (Architecture) ۴۷
۱.۲ کدگذاری موقعیت (Positional Encoding) ۴۷
۲.۲ بردارکننده (Tokenizer) ۴۷
۳.۲ شبکه عصبی (MLP) ۴۷
۴.۲ تابع فعالساز (Activation Function) ۴۸
۵.۲ نرمالسازی (Normalization) ۴۸
۶.۲ مکانیزم توجه (Attention) ۴۸
۷.۲ ساختار ترنسفورمر (Transformer) ۴۹

۳ بهینه سازی (Optimization) ۴۹
۱.۳ الگوریتم (Algorithm) ۴۹
۲.۳ تابع هزینه (Loss Function) ۵۰

۴ تنظیم هایپرپارامتر (Hyperparameter Tuning) ۵۰

۵ قوانین مقیاس پذیری (Scaling Laws) ۵۰
۱.۵ مقدمه ۵۰
۲.۵ اصلاح و توسعه قوانین مقیاس‌پذیری کلاسیک ۵۱
۳.۵ تاثیر معماری و شکل مدل (Model Shape & Architecture) ۵۱
۴.۵ قوانین مقیاس‌پذیری برای مدل‌های ترکیبی از خبرگان (MoE) ۵۲
۵.۵ نقش داده‌ها: کیفیت، ترکیب و چگالی ۵۲
۶.۵ فراتر از خطای مطلق: معیارهای نوین ارزیابی ۵۲
۷.۵ نتیجه‌گیری ۵۳

۶ ارزیابی (Evaluation) ۵۳

۷ یادگیری تقویتی (Reinforcement Learning) ۵۳
۱.۷ الگوریتم بهینه‌سازی تقریبی خط‌مشی (Proximal Policy Optimization - PPO) ۵۸
۲.۷ الگوریتم بهینه‌سازی نسبی خط‌مشی گروهی (Group Relative Policy Optimization - GRPO) ۶۵
۳.۷ الگوریتم بهینه‌سازی توالی گروهی خط‌مشی (Group Sequence Policy Optimization - GSPO) ۷۰
۴.۷ الگوریتم بهینه‌سازی خط‌مشی با برش مجزا و نمونه‌برداری پویا (Decoupled Clip and Dy-namic Sampling Policy Optimization - DAPO) ۷۷
۵.۷ الگوریتم بهینه‌سازی خط‌مشی تطبیقی نرم (Soft Adaptive Policy Optimization - SAPO) ۸۳
۶.۷ الگوریتم بهینه‌سازی خط‌مشی با ضریب اهمیت برش‌خورده (Clipped IS-weight Policy Optimization - CISPO) ۸۹
۷.۷ الگوریتم یادگیری تقویتی با پاداش‌های اعتبارسنجی‌پذیر (Reinforcement Learning with Verifiable Rewards - RLVR) ۹۵
۸.۷ الگوریتم یادگیری تقویتی با پاداش‌های قابل اعتبارسنجی و بهینه‌سازی گروهی (OlmoRL: RLVR + GRPO) ۱۰۰

۸ منابع **۱۰۸**

۱ آموزش (Training)

۱.۱ داده‌ها (Data)

۱.۱.۱ پیش‌آموزش (Pre-Training)

- **DeepSeek-V4**

مقدمه و ساختار کلی دادگان پیش‌آموزش فاز پیش‌آموزش (Pre-training) در مدل‌های زبانی سری **DeepSeek-V4** با هدف شکستن سد محاسباتی در پردازش توالی‌های فوق‌طولانی تا سقف بافت یک میلیون توکن (1M-token context) بازطراحی شده است. در این چارچوب، دو نسخه پایه با مقیاس‌های محاسباتی متمایز آموزش دیده‌اند:

- **مدل DeepSeek-V4-Flash:** دارای ۲۸۴ میلیارد پارامتر کل (۱۳ میلیارد پارامتر فعال در هر توکن) که روی مجموعه‌ای بالغ بر **۳۲ تریلیون توکن** (32T tokens) پیش‌آموزش دیده است.
- **مدل DeepSeek-V4-Pro:** دارای ۶.۱ تریلیون پارامتر کل (۴۹ میلیارد پارامتر فعال در هر توکن) که فرایند پیش‌آموزش آن بر روی بیش از **۳۳ تریلیون توکن** (33T tokens) صورت پذیرفته است.

هدف بهینه‌سازی پایه در طول این فرایند، حداقل‌سازی تلفات خودرگرسیو مدل‌سازی زبان (Autoregressive Language Modeling Loss) بر روی توالی توکن‌ها است:

$$\mathcal{L}_{\mathsf{LM}}(\theta) = -\sum_{t=1}^{N} \log P_\theta(x_t \mid x_1, x_2, \ldots, x_{t-1}) \tag{۱}$$

ترکیب‌بندی و پالایش پیکره داده‌ها (Data Construction) پیکره متنی پیش‌آموزش بر پایه تجربیات نسخه DeepSeek-V3 [۱] شکل گرفته، اما با هدف ارتقای چگالی دانش، جلوگیری از افت کیفی و پشتیبانی از استدلال عمیق، اصلاحات اساسی در حوزه‌های زیر اعمال شده است:

- **پالایش وب و پیشگیری از فروپاشی مدل (Mitigating Model Collapse):** به منظور مقابله با پدیده فروپاشی مدل ناشی از انباشت داده‌های بازتولیدشده توسط هوش مصنوعی و متون قالبی در وب باز [۲]، خط لوله‌های فیلترینگ سخت‌گیرانه‌ای تعبیه شد تا محتواهای ماشینی، تکراری و قالبی به صورت دسته‌ای شناسایی و پالایش شوند.
- **تزریق داده‌های عامل‌محور در آموزش میانی (Agentic Mid-Training Data):** پیکره ریاضیات و برنامه‌نویسی به عنوان هسته اصلی تفکر مدل حفظ شده است؛ علاوه بر آن، در فاز میانی آموزش (Mid-Training Phase) داده‌های تعاملی و ابزارمحور (Agentic data)—شامل ردپای خط فرمان، محیط‌های چندمرحله‌ای رفع باگ و بازخورد مفسرها—به مدل تزریق شد تا بازنمایی کد از متن ایستا به کنشگری پویا ارتقا یابد.
- **توسعه داده‌های چندزبانه و دانش حاشیه‌ای (Long-Tail Knowledge):** برای بازنمایی دقیق فرهنگ‌ها و زبان‌های گوناگون، دامنه دادگان چندزبانه توسعه یافت تا پوشش مدل در بازیابی دانش حاشیه‌ای با فراوانی آماری اندک تضمین گردد.
- **تمرکز راهبردی بر اسناد طولانی دانشگاهی (Long-Document Curation):** به منظور بارورسازی درک زمینه در مقیاس‌های فراتر از چند صدهزار کلمه، مقالات پژوهشی، گزارش‌های تخصصی و اسناد فنی دارای پیوستگی عمیق معنایی با اولویت بالا گزینش شدند.

پیش‌پردازش، توکنایزیشن و مهندسی نمونه‌ها (Data Pre-Processing) فرایند مهندسی داده‌ها جهت بهینه‌سازی ورود توکن‌ها به سازوکار توجه شامل راهکارهای زیر است:

- **واژگان و نشانه‌گذار (Tokenizer):** دایره واژگان در اندازه 128K توکن تثبیت شده و نشانه‌های ویژه‌ای (Special tokens) جهت ساختاربندی و مدیریت بافت‌های متنی طولانی افزوده شده است.
- **تکمیل درون‌متنی (Fill-in-the-Middle - FIM):** جهت تعمیق درک دوطرفه متن، راهبرد FIM [۱] اعمال گردید که طی آن سند متنی به سه بخش پیشوند (P)، میانه (M) و پسوند (S) تفکیک و به فرم زیر بازآرایی می‌شود:

$$\mathcal{D}_{\text{FIM}} = [\text{<PRE>}] \circ P \circ [\text{<SUF>}] \circ S \circ [\text{<MID>}] \circ M \tag{۲}$$

این رویکرد مدل را قادر می‌سازد با شرط‌گذاری هم‌زمان روی ابتدا و انتهای بافت، بخش میانی را پیش‌بینی کند.

- **بسته‌بندی اسناد بدون اتلاف توکن (Document Packing):** با اقتباس از رهیافت دینگ و همکاران [۳]، به‌جای قطع نامتقارن (Truncation) متون در انتهای پنجره بافت، اسناد برگرفته از منابع گوناگون بر پایه طول دسته‌بندی و در قالب بسته‌های بهینه در کنار یکدیگر قرار گرفتند تا کمترین برش رخ دهد.
- **ماسک توجه در سطح نمونه (Sample-Level Attention Masking):** برخلاف DeepSeek-V3، در این نسخه ماسک‌گذاری توجه در سطح سند فعال شد. ماتریس توجه M به‌گونه‌ای پیکربندی می‌شود که از سرایت توجه میان اسناد مستقلی که در یک توالی بسته‌بندی شده‌اند جلوگیری شود:

$$M_{ij} = \begin{cases} 0, & \text{if} i \geq j \text{ and } \text{DocID}(i) = \text{DocID}(j) \\ -\infty, & \text{otherwise} \end{cases} \tag{۳}$$

برنامه درسی طول بافت و پویایی بهینه‌سازی (Context Curriculum & Optimization) فرایند آموزش مدل روی بافت‌های میلیونی به شیوه گام‌به‌گام و مرحله‌ای (Curriculum Learning) مدیریت شده است:

- **توالی افزایش بافت:** آموزش از طول توالی 4K آغاز شده و به توالی‌های 16K، سپس 64K و در نهایت 1M توکن تعمیم یافته است.
- **مقیاس دسته‌ها (Batch Size Scheduling):** در نسخه Flash اندازه دسته تا سقف 75.5M توکن و در نسخه Pro تا مرز 94.4M توکن در هر گام افزایش یافته است.
- **ورود تدریجی توجه تنک:** در طول ۱ تریلیون توکن نخست، مدل با سازوکار توجه متراکم (Dense Attention) گرم‌سازی شد. سپس در طول بافت 64K، معماری توجه تنک فشرده (CSA) با یک مرحله پایدارسازی کوتاه‌مدت برای شاخص‌گذار Lightning Indexer فعال گردید و تا پایان فاز پیش‌آموزش ادامه یافت.
- **نرخ یادگیری:** مدل‌ها با بهینه‌ساز Muon [۴، ۵] و تنظیمات نرخ یادگیری خطی اولیه و واپاشی کسینوسی نهایی (از 2.7×10^{-4} به 2.7×10^{-5} در مدل Flash و از 2.0×10^{-4} به 2.0×10^{-5} در مدل Pro) به پیش‌آموزش رسانده شدند.

مهار ناپایداری‌های ناشی از توزیع داده‌ها جهت رفع جهش‌های زیان (Loss Spikes) ناشی از داده‌های پرت در لایه‌های MoE، دو تکنیک در خط لوله داده و محاسبات اعمال گردید:

- **مسیریابی پیش‌دستانه (Anticipatory Routing):** شاخص‌های گیتینگ بر اساس وزن‌های با تأخیر تاریخی $\theta_{t-\Delta t}$ روی داده‌های پیش‌واکشی‌شده محاسبه شده تا چرخه هم‌افزایی مثبت مقادیر فرین شکسته شود.
- **تحدید دامنه SwiGLU:** ورودی‌های خطی فعال‌ساز SwiGLU [۶] به بازه $[-10, 10]$ و مؤلفه گیت آن به حداکثر مقدار 10 تحدید شدند.

ارزیابی‌های آماری مدل‌های پایه (Base Model Evaluations) نتایج ارزیابی مدل‌های پایه پیش‌آموزش‌دیده (پیش از هرگونه هم‌ترازسازی یا Post-training) در جدول زیر گزارش شده است. این نتایج نشان می‌دهد که مدل DeepSeek-V4-Flash-Base با وجود دارا بودن تنها ۱۳ میلیارد پارامتر فعال، به برکت کیفیت داده‌ها و راهبردهای پیش‌پردازش، مدل پیشین DeepSeek-V3.2-Base (با ۳۷ میلیارد پارامتر فعال) را در بنچمارک‌های متعددی نظیر C-Eval (92.1 در برابر 90.4)، MultiLoKo (42.2 در برابر 38.7) و LongBench-V2 (44.7 در برابر 40.2) پشت سر گذاشته است. همچنین مدل DeepSeek-V4-Pro-Base در آزمون حقیقت‌سنجی دادگان حاشیه‌ای (FACTS Parametric) به جهش چشمگیر 62.6 (رشد نسبی ۱۳۱٪ نسبت به 27.1) دست یافته است.

جدول ۱: ارزیابی جامع عملکرد مدل‌های پایه **DeepSeek-V4** در مقایسه با **DeepSeek-V3.2-Base**

دسته ارزیابی	بنچمارک (متریک)	پرومپت (Shots)	DeepSeek-V3.2-Base	DeepSeek-V4-Flash-Base	DeepSeek-V4-Pro-Base
مشخصات	معماری	-	MoE	MoE	MoE
	پارامترهای فعال	-	۳۷B	۱۳B	۴۹B
	پارامترهای کل	-	۶۷۱B	۲۸۴B	۶T.۱
دانش جهان (World Knowledge)	AGIEval (EM)	۰-shot	۱.۸۰	۶.۸۲	**۱.۸۳**
	MMLU (EM)	۵-shot	۸.۸۷	۷.۸۸	**۱.۹۰**
	MMLU-Redux (EM)	۵-shot	۵.۸۷	۴.۸۹	**۸.۹۰**
	MMLU-Pro (EM)	۵-shot	۵.۶۵	۳.۶۸	**۵.۷۳**
	MMMLU (EM)	۵-shot	۹.۸۷	۸.۸۸	**۳.۹۰**
	C-Eval (EM)	۵-shot	۴.۹۰	۱.۹۲	**۱.۹۳**
	CMMLU (EM)	۵-shot	۹.۸۸	۴.۹۰	**۸.۹۰**
	MultiLoKo (EM)	۵-shot	۷.۳۸	۲.۴۲	**۱.۵۱**
	Simple-QA verified (EM)	۲۵-shot	۳.۲۸	۱.۳۰	**۲.۵۵**
	SuperGPQA (EM)	۵-shot	۰.۴۵	۵.۴۶	**۹.۵۳**
	FACTS Parametric (EM)	۲۵-shot	۱.۲۷	۹.۳۳	**۶.۶۲**
	TriviaQA (EM)	۵-shot	۳.۸۳	۸.۸۲	**۶.۸۵**
زبان و استدلال (Lang. & Reas.)	BBH (EM)	۳-shot	**۶.۸۷**	۹.۸۶	۵.۸۷
	DROP (F1)	۱-shot	۲.۸۸	۶.۸۸	**۷.۸۸**
	HellaSwag (EM)	۰-shot	۴.۸۶	۷.۸۵	**۰.۸۸**
	WinoGrande (EM)	۰-shot	۹.۷۸	۵.۷۹	**۵.۸۱**
	CLUEWSC (EM)	۵-shot	۵.۸۳	۲.۸۲	**۲.۸۵**
کد و ریاضیات (Code & Math)	BigCodeBench (Pass@1)	۳-shot	**۹.۶۳**	۸.۵۶	۲.۵۹
	HumanEval (Pass@1)	۰-shot	۸.۶۲	۵.۶۹	**۸.۷۶**
	GSM8K (EM)	۸-shot	۱.۹۱	۸.۹۰	**۶.۹۲**
	MATH (EM)	۴-shot	۵.۶۰	۴.۵۷	**۵.۶۴**
	MGSM (EM)	۸-shot	۳.۸۱	**۷.۸۵**	۴.۸۴
	CMath (EM)	۳-shot	۶.۹۲	**۶.۹۳**	۹.۹۰
بافت طولانی (Long Context)	LongBench-V2 (EM)	۱-shot	۲.۴۰	۷.۴۴	**۵.۵۱**

- **GLM-5**

مقدمه و ساختار کلان دادگان پایه (Foundational Overview & Token Budget) مدل زبانی **GLM-5** [۷] با هدف گذار از پارادایم کدنویسی حسی و تعاملی ساده (Vibe Coding) به سوی مهندسی عاملی خودکار و صنعتی (Agentic Engineering) بازطراحی شده است. در این راستا، مدل پایه (Base Model) بر روی پیکره عظیمی به حجم مجموع **۵.۲۸ تریلیون توکن** (28.5T tokens) آموزش داده شده است. معماری مدل پایه دارای ۷۴۴ میلیارد پارامتر کل و ۴۰ میلیارد پارامتر فعال در هر توکن است که در قالب ساختار تخصص‌یافته با ۲۵۶ متخصص (MoE) پیکربندی شده است.

فرایند تغذیه داده در مدل پایه **GLM-5** به شکل یک خط لوله سه‌مرحله‌ای پیوسته مهندسی شده است:

$$\mathcal{D}_{\text{Base}} = \mathcal{D}_{\text{Pre-training}} \cup \mathcal{D}_{\text{Mid-training}} \cup \mathcal{D}_{\text{DSA-Adaptation}} \tag{۴}$$

توزیع بودجه توکن‌ها در طول پنجره‌های بافت (Context Window) به شرح زیر است:

- **پیش‌آموزش اولیه (General Pre-training):** شامل **۲۷ تریلیون توکن** با طول بافت استاندارد 4K، تفکیک‌شده به ۱۸ تریلیون توکن متون عمومی و چندزبانه و ۹ تریلیون توکن کدهای برنامه‌نویسی و استدلال ریاضی اولیه.
- **آموزش میانی (Mid-training):** شامل **۵۵.۱ تریلیون توکن** جهت آماده‌سازی شناختی بافت بلند و داده‌های عاملی در سه پله متوالی (32K با حجم 1T، 128K با حجم 500B و 200K با حجم 50B).

- **تطبیق توجه تنک دیپ‌سیک (DSA Adaptation):** شامل **۲۰ میلیارد توکن** با طول بافت 200K در قالب پیش‌آموزش پیوسته (Continued Pre-Training).

ترکیب‌بندی و پالایش پیکره پیش‌آموزش اولیه (Pre-Training Data Domains) پیکره متنی فاز پیش‌آموزش اولیه به منظور ارتقای چگالی حقیقت‌سنجی و مهارت در استدلال، دستخوش دگرگونی‌های ساختاری در سه حوزه کلیدی شده است:

- **پالایش وب و استخراج دانش حاشیه‌ای (Web Corpus & Long-tail Knowledge):** افزون بر طبقه‌بندهای آماری مرسوم، یک طبقه‌بند برداری مبتنی بر تعبیه معنایی جملات بر پایه چارچوب DCLM [۸] آموزش داده شد تا متون غنی از لحاظ مفهومی از وب باز استخراج گردند. همچنین با هدف رفع خلأ دانش دم‌بلند (Long-tail Knowledge)، «طبقه‌بند دانش جهان» (World Knowledge Classifier) بهینه‌سازی‌شده با مدخل‌های ویکی‌پدیا و برچسب‌گذاری‌های مدل زبانی تعبیه گردید تا اطلاعات دانشنامه‌ای ارزشمند از صفحات وب با کیفیت متوسط بازیابی شوند.
- **غنی‌سازی پیکره کد و حذف فازی تکرارها (Code Corpus & Fuzzy Deduplication):** مخزن کدهای پیش‌آموزش با برداشت نسخه‌های تازه از پلتفرم‌های میزبانی کد و صفحات وب حاوی کد به‌روزرسانی شد. با اجرای الگوریتم‌های حذف فازی موارد تکراری (Fuzzy Deduplication)، حجم توکن‌های یکتای کد به میزان **۲۸٪** افزایش یافت:

$$\Delta N_{\text{unique}} = \frac{N_{\text{unique}}^{\text{GLM-۵}} - N_{\text{unique}}^{\text{GLM-۵.۴}}}{N_{\text{unique}}^{\text{GLM-۵.۴}}} = +28\% \qquad (۵)$$

  علاوه بر این، ناهماهنگی‌های فراداده‌ای در مخزن Software Heritage تصحیح و طبقه‌بندهای اختصاصی برای زبان‌های برنامه‌نویسی کم‌منبع (Low-resource Languages) نظیر Scala، Swift و Lua جهت نمونه‌برداری آگاه از کیفیت پیاده‌سازی شدند.
- **داده‌های ریاضیات، علوم و پالایش متون ساختگی (Math & Science Curation):** لوله‌های استخراج محتوای وب و پارس فایل‌های چندستونی PDF مقالات علمی و کتاب‌های ریاضی بهبود یافتند. در اسناد و کتب حجیم، یک الگوریتم اختصاصی «تکه‌تکه کردن و تجمیع نمره» (Chunk-and-Aggregate Scoring) به کار رفت تا نمره غنای آموزشی اسناد به درستی برآورد شود. به‌منظور جلوگیری از سقوط مدل (Model Collapse) و حفظ اصالت استدلال، خطوط لوله فیلترینگ به گونه‌ای تنظیم شدند که استفاده از هرگونه داده سنتزی، تولیدشده با هوش مصنوعی یا داده‌های الگومحور (Template-based) در فاز پیش‌آموزش ریاضیات اکیداً حذف شود.

دادگان آموزش میانی و برنامه درسی طول بافت (Mid-Training & Context Scaling) برای تثبیت پایداری مدل در پردازش کدبیس‌های عظیم، اسناد فوق‌طولانی و جریان‌های کاری عاملی، مرحله آموزش میانی با بسط تصاعدی طول بافت انجام پذیرفت:

- **توالی بسط بافت (Progressive Context Expansion):** بافت مدل از طول پایه 4K طی سه گام پیاپی شامل مرحله 32K با ۱ تریلیون توکن، مرحله 128K با ۵۰۰ میلیارد توکن و نهایتاً مرحله 200K با ۵۰ میلیارد توکن ارتقا یافت.
- **داده‌های مهندسی نرم‌افزار در سطح مخزن (Software Engineering Data):** مدل با الحاق ساختار کامل مخزن گیت‌هاب، تاریخچه تغییرات (Commit Diffs)، گزارش باگ‌ها (Issues)، درخواست‌های ادغام (Pull Requests) و فایل‌های مرتبط به صورت یک توالی پیوسته آموزش دید. با تعدیل ضوابط فیلترینگ در سطح مخزن و تشدید نظارت کیفی در سطح تک‌تک Issueها، بالغ بر **۱۰ میلیون جفت Issue–PR** استخراج شد که پیکره نهایی آن پس از پالایش شامل **۱۶۰ میلیارد توکن یکتا** (160B unique tokens) گردید.
- **مهار پدیده گم‌شدن در میانه (Mitigating Lost-in-the-Middle):** با الهام از ساختارهای NextLong [۹] و EntropyLong [۱۰]، داده‌های سنتزی برای بازتولید وابستگی‌های دوربرد طراحی شدند. متون به شدت مشابه از طریق بسته‌بندی درهم‌تنیده (Interleaved Packing) ترکیب شدند تا مدل ناچار به جست‌وجو در میانه بافت شود. در پله 200K، داده‌های شبیه به MRCR [۱۱] با تنوع بالا جهت تثبیت یادآوری در دیالوگ‌های طولانی تزریق شدند.

داده‌های تطبیق توجه تنک دیپ‌سیک (DSA Continued Pre-Training Data) به جای آموزش پرهزینه توجه از مبدا، مدل متراکم انتهای فاز آموزش میانی از طریق استراتژی دونشستی «گرم‌کردن متراکم و تطبیق آموزش تنک» به سازوکار DeepSeek Sparse Attention (DSA) [۱۲] مجهز گردید:

- **مرحله گرم‌کردن نمایه‌ساز (Indexer Warm-up):** شاخص‌گذار Lightning Indexer به مدت ۱،۰۰۰ گام آموزش دید. در هر گام ۱۴ توالی به طول ۲۰۲،۷۵۲ توکن (مجموعاً ۸۴.۲ میلیارد توکن) با نرخ یادگیری بیشینه 5×10^{-3} به مدل ارائه شد در حالی که وزن‌های پایه ثابت بودند.
- **مرحله تطبیق تنک (Sparse Adaptation Stage):** آموزش پیوسته کل مدل بر روی **۲۰ میلیارد توکن** از توزیع داده‌های فاز آموزش میانی با نرخ یادگیری ثابت 1×10^{-5} صورت پذیرفت. این استراتژی کم‌هزینه توانست عملکرد توجه تنک را بدون اتلاف کارایی بافت بلند، با مدل متراکم اولیه معادل سازد.

ارزیابی‌های آماری مدل‌های پایه (Base Model Evaluations) نتایج ارزیابی مدل پایه پیش‌آموزش‌دیده **GLM-5-Base** (پیش از ورود به فازهای SFT و RL) در جدول زیر ارائه شده است. در آزمون‌های ارزیابی کد، مدل به برکت تزریق داده‌های یکتای مهندسی نرم‌افزار به امتیاز خیره‌کننده 87.0 در بنچمارک EvalPlus دست یافته است (رشد نسبی ۴٪.۱۱ نسبت به **GLM-4.5-Base** و ۶٪.۳۲ نسبت به **DeepSeek-V3-Base**). همچنین در آزمون حقیقت‌سنجی SimpleQA، دستاورد امتیاز 36.0 کارآمدی طبقه‌بندهای دانش وب را اثبات می‌کند. افت مشاهده‌شده در نمرات خام بنچمارک‌های ریاضیاتی نظیر GSM8K، پیامد مستقیم تصمیم آگاهانه تیم توسعه مبنی بر حذف داده‌های سنتزی و قالبی در پیش‌آموزش بوده که این نقص متعاقباً در فاز یادگیری تقویتی استدلالی (Reasoning RL) با کسب نمره برتر 50.4 در HLE جبران گردیده است.

جدول ۲: ارزیابی جامع عملکرد مدل پایه **GLM-5-Base** در مقایسه با سایر مدل‌های پایه شاخص

دسته ارزیابی	بنچمارک (متریک)	DeepSeek-V3-Base	Kimi-K2-Base	GLM-4.5-Base	GLM-5-Base
مشخصات	معماری	MoE	MoE	MoE	MoE
	پارامترهای فعال	۳۷B	۳۲B	۳۲B	**۴۰B**
	پارامترهای کل	۶۷۱B	۱۰۴۳B	۳۵۵B	**۷۴۴B**
انگلیسی (English)	SimpleQA (EM)	۶.۲۶	۳.۳۵	۰.۳۰	**۰.۳۶**
	BBH (EM)	۴.۸۸	**۷.۸۸**	۲.۸۶	۴.۸۷
	MMLU (EM)	۲.۸۷	۸.۸۷	۱.۸۶	**۳.۸۸**
	HellaSwag (EM)	۹.۸۸	**۶.۹۴**	۱.۸۷	۱.۸۸
	PIQA (EM)	۷.۸۴	–	**۳.۸۵**	۶.۸۴
	TriviaQA (EM)	۹.۸۲	**۱.۸۵**	۰.۸۰	۹.۸۰
کدنویسی (Code)	EvalPlus (Pass@1)	۶.۶۵	۳.۸۰	۱.۷۸	**۰.۸۷**
	LiveCodeBench-Base (Pass@1)	۶.۲۴	۳.۲۶	۱.۲۸	**۴.۳۴**
ریاضیات (Math)	GSM8K (EM)	۶.۸۷	**۱.۹۲**	۴.۷۹	۸.۶۸
	MATH (EM)	۶.۶۲	**۲.۷۰**	۰.۶۱	۴.۵۶
چینی (Chinese)	CLUEWSC (EM)	۷.۸۲	–	۵.۸۳	**۲.۸۴**
	C-Eval (EM)	۱.۹۰	**۵.۹۲**	۹.۸۶	۸.۸۸
	C3 (EM)	۶.۷۸	–	**۱.۸۳**	۳.۸۰
	Chinese-SimpleQA (EM)	۱.۷۲	**۶.۷۷**	۱.۷۰	۶.۷۴

- **Kimi K3**

مقدمه و فلسفه دادگان پیش‌آموزش چندوجهی بومی فاز پیش‌آموزش (Pre-training) مدل **Kimi K3** با هدف شکستن مرزهای محاسباتی در مقیاس‌های فراتریلیونی و دستیابی به بافتار فوق‌طولانی ۱ میلیون توکن (1M-token context) بر پایه معماری تنک ترکیبی از متخصصان (Native Multimodal MoE) با ۷۸.۲ تریلیون پارامتر کل و ۲.۱۰۴ میلیارد پارامتر فعال طراحی شده است [۱۳]. برخلاف رویکردهای رایج که از یک مرحله انطباق پسینی (Post-hoc Modality Alignment) جهت اتصال برج بینایی منجمد به مدل زبانی متنی استفاده می‌کنند، در **Kimi K3** یک راهبرد یادگیری چندوجهی بومی

اتخاذ شده است که در آن توکن‌های متنی و بصری از همان گام نخست آموزش به‌صورت درهم‌تنیده تحت هدف واحد پیش‌بینی توکن بعدی (Next-Token Prediction) بهینه‌سازی می‌شوند:

$$\mathcal{L}_{\text{pretrain}}(\theta) = -\sum_{t=1}^{N} \log P_\theta(u_t \mid u_1, u_2, \ldots, u_{t-1}) \tag{۶}$$

که در آن u_t نمایانگر یک توکن متنی یا یک توکن بصری فشرده‌شده در فضای برداری مشترک است. ارتقای هم‌زمان خط‌لوله داده، پویایی بهینه‌سازی و بازنویسی متون موجب دستیابی به **بهره‌وری مقیاس‌بندی ۵.۲ برابری** (Scaling Efficiency gain $2.5\times$) در مقایسه با نسل پیشین یعنی **Kimi K2** [۱۴] روی داده‌های اعتبارسنجی خارج از توزیع (Held-out OOD) شده است.

ترکیب‌بندی و خط لوله پالایش پیکره متنی (Textual Corpus Curation) پیکره متنی مدل بر چهار دامنه بنیادین استوار است که با بهره‌گیری از زیرساخت پیشرفته پالایش [۱۴، ۱۵] پردازش شده‌اند:

- **متن وب (Web Text):** پوشش‌دهنده تنوع زبانی عمومی و دانش عام حاکم بر بستر اینترنت.
- **کد (Code):** مخازن سورس‌کد چندزبانه، مستندات فنی و ساختارهای الگوریتمی جهت تقویت توان استدلال صوری.
- **ریاضیات (Mathematics):** اسناد تحلیلی، متون دارای نمادگذاری علمی لاتک (LaTeX) و اثبات‌های ریاضیاتی.
- **دانش تخصصی (Knowledge):** دانشنامه‌ها، کتب مرجع، اسناد حقوقی، داده‌های مالی و مقالات دانشگاهی بازبینی‌شده.

فرایند پاک‌سازی داده‌ها از سه گام دقیق عبور می‌کند:

۱. **فیلترهای اکتشافی مبتنی بر قاعده (Rule-based Heuristics):** حذف محتوای باینری، کدهای خراب، لاگ‌های سیستمی بی‌معنی و متون نامتعارف ماشینی.
۲. **امتیازدهی کیفی با طبقه‌بند (Classifier-based Quality Scoring):** تخمین نمره کیفیت اسناد و غربال نمونه‌های با چگالی اطلاعاتی نازل.
۳. **حذف افزونگی چندمرحله‌ای (Deduplication):** تلفیق حذف تکرار دقیق با نگاشت‌های قطعی (Exact Dedup) و حذف تکرار فازی (Fuzzy Dedup) با تحلیل شباهت سطحی اسناد.
۴. **نرخ‌های نمونه‌برداری بهینه (Sampling Rates):** تخصیص بودجه و وزن به هر دامنه بر اساس مطالعات تجربی گسترده بر روی مدل‌های کوچکتری از همان خانواده صورت پذیرفته است.

دستورالعمل بازنویسی داده‌ها با اعتبارسنجی وفاداری (Fidelity-Verified Rephrasing) به منظور برطرف ساختن اطناب کلام و نویزهای ساختاری در متون وب، رویه بازنویسی معرفی‌شده در **Kimi K2** [۱۴] با سه راهبرد محوری توسعه یافته است:

- **پرامپت‌نویسی با تنوع دیدگاه و سبک (Style and Perspective-Diverse Prompting):** متون تخصصی ریاضیاتی و دانشی با زاویه دیدهای متنوع تحلیلی و آموزشی بازآفرینی می‌شوند.
- **تولید خودهمبسته تکه‌ای (Chunk-wise Autoregressive Generation):** فرایند بازنویسی متن در قالب قطعات منسجم صورت می‌گیرد تا یکدستی ساختار استدلالی تضمین شود.
- **راستی‌آزمایی وفاداری به منبع (Fidelity Verification):** تمامی داده‌های بازنویسی‌شده توسط ممیزهای خودکار ارزیابی می‌شوند تا با انطباق با سند مرجع، از ورود هرگونه توهم یا داده ساختگی به دادگان پیش‌آموزش جلوگیری شود:

$$\text{Fidelity}(D_{\text{rephrased}}, D_{\text{source}}) = \mathbb{I}\left(\text{FactCheck}(D_{\text{rephrased}}, D_{\text{source}}) \geq \tau_{\text{fidelity}}\right) \tag{۷}$$

پیکره چندوجهی و آموزش رمزگذار بصری از صفر (MoonViT-V2 from Scratch) برخلاف سنت جاری در مدل‌های چندوجهی که رمزگذار تصویر را بر پایه مدل‌های کنتراستیو نظیر SigLIP مقداردهی اولیه می‌کنند، رمزگذار بینایی **Kimi K3** تحت عنوان **MoonViT-V2** (با ۴۰۱ میلیون پارامتر و ۲۷ لایه ViT) به‌طور کامل **از صفر (From Scratch)** تحت تابع هدف پیش‌بینی توکن بعدی آموزش می‌بیند:

- **پایداری بهینه‌سازی و مهار جهش‌های گرادیان:** آموزش توأم رمزگذارهای کنتراستیو با مدل‌های زبانی عظیم، تلاطم‌های شدیدی در نُرم گرادیان برج بینایی پدید می‌آورد. همان‌طور که در شواهد گزارش فنی مشخص است، مدل پیشین با مقداردهی SigLIP دچار جهش‌های مقطعی مکرر در نُرم گرادیان می‌شد، درحالی‌که **MoonViT-V2** نُرم‌های گرادیان بسیار پایین‌تر و کاملاً پایداری را تجربه می‌کند.
- **داده‌های برنامه‌نویسی بصری (Programmatic Multimodal Data):** علاوه بر داده‌های کلاسیک تصویر-زیرنویس، اسناد متنی-تصویری درهم‌تنیده و OCR، حجم عظیمی از داده‌های برنامه‌نویسی شامل کدهای منبع به همراه خروجی‌های رندرشده دوبعدی و سه‌بعدی (SVG، صفحات وب، بازی‌های رایانه‌ای، دارایی‌های سه‌بعدی و نقشه‌های مهندسی CAD) به پیکره آموزشی تزریق گردید.
- **نظارت مختصات فضایی:** تمام نمونه‌های نیازمند درک مکان به دو صورت مختصات مطلق پیکسلی و نرمال‌شده در بازه $[0,1]^2$ برچسب‌گذاری شدند.
- **فشرده‌سازی پیکسلی توکن‌ها (Pixel-Shuffle Downsampling):** تصاویر و فریم‌های ویدئویی تا وضوح بالای 3584×3584 با کاهش نمونه‌برداری 2×2 از طریق عملگر Pixel-Shuffle متراکم شده و تعداد توکن‌های بینایی را با ضریب ۴ کاهش می‌دهند تا گنجایش بافت ۱ میلیون توکنی حفظ شود:

$$T_{\text{visual}} = \frac{1}{4} \cdot \left(\frac{H_{\text{img}}}{P} \times \frac{W_{\text{img}}}{P} \right), \quad P = 14 \tag{۸}$$

مهندسی داده‌های بافتار فوق‌بلند و مسائل مصنوعی به منظور تثبیت ظرفیت بازیابی و استدلال در سقف بافت ۱ میلیون توکن:

- **پالایش ویدئوها با هش ادراکی (Perceptual Hashing):** فریم‌های ویدئویی تکراری و ساکن با استفاده از pHash غربال می‌شوند تا از افزونگی بیهوده توکن‌های بصری جلوگیری گردد.
- **وزن‌دهی افزایشی در فاز سردسازی (Upsampling during Cooldown):** به سبب کمیابی طبیعی اسناد طولانی منسجم در وب، فراوانی اسناد طولانی و ویدئوها در مراحل پایانی پیش‌آموزش و به‌ویژه در فاز سردسازی (Cooldown Phase) به‌طور چشمگیری تقویت گردید.
- **تولید وظایف ترکیبی با وابستگی سراسری (Synthetic Interleaved Tasks):** برای جلوگیری از فروپاشی سازوکارهای توجه به الگوهای محلی، اسناد و زیروظایف گوناگون به‌گونه‌ای چیده و ادغام شدند که حل مسئله نهایی منوط به تلفیق اطلاعاتی باشد که در سراسر بازه ۱ میلیون توکن پراکنده‌اند:

$$y_T = f(x_{t_1}, x_{t_2}, \ldots, x_{t_k}), \quad 0 \ll t_1 \ll t_2 \ll \cdots \ll t_k \leq 10^6 \tag{۹}$$

برنامه درسی پیشرونده طول بافتار (Progressive Context Curriculum) مدیریت محاسبات پیش‌آموزش بر روی طول توالی‌های بالا از طریق یک برنامه درسی چهارمرحله‌ای انجام گرفته است:

- **مرحله اول (پیش‌آموزش پایه):** آموزش در طول بافت 8K توکن آغاز می‌شود.
- **مرحله دوم (گسترش میانی):** افزایش پنجره بافتار به 64K توکن در ادامه فرایند پیش‌آموزش.
- **مرحله سوم و چهارم (فاز سردسازی):** جهش متوالی طول بافت از 256K به 1M توکن در طول دوره Cooldown.

با اتکا به مکانیزم زوال ذاتی لایه‌های توجه خطی KDA [۱۶] و حذف کامل کدهای موضعی صریح (NoPE) در لایه‌های Gated MLA، مدل بدون نیاز به دستکاری پایه‌های فرکانسی RoPE یا روش‌های برون‌یابی نظیر YaRN [۱۷]، مستقیماً به بافتار ۱ میلیون توکنی تعمیم می‌یابد.

قوانین مقیاس‌پذیری داده‌ها و برنامه واپاشی نرخ یادگیری روابط تجربی حاکم بر مصرف داده و توان پردازشی بر پایه قوانین توان (Power-law Scaling) تدوین شده‌اند:

$$\mathcal{L}_{\text{val}}(C) = \mathcal{L}_{\infty} + \left(\frac{C_0}{C}\right)^{\alpha} \tag{۱۰}$$

برازش منحنی‌های قانون مقیاس‌پذیری حاکی از آن است که **Kimi K3** برای دستیابی به تراز زیان اعتبارسنجی مشابه با **Kimi K2**، به تقریب تنها به ۴۰٪ از حجم پردازش ($C_{\text{K۳}} \approx 0.4 \cdot C_{\text{K۲}}$) نیاز دارد.

علاوه بر این، در مقایسه زمان‌بند واپاشی کسینوسی (Cosine Decay) با زمان‌بند گرمایش-پایداری-زوال (WSD) [۱۸]، نشان داده شد که با اجرای مستقل جستجوی قوانین مقیاس‌پذیری جهت تعیین نرخ یادگیری اوج و اندازه دسته بهینه برای هر روش، رویکرد **واپاشی کسینوسی با ۱٪ گرمایش خطی** به طور مداوم زیان نهایی پایین‌تری نسبت به WSD ثبت کرده و به عنوان زمان‌بند رسمی پیش‌آموزش تثبیت شد. پارامترهای ماتریسی مدل نیز با بهینه‌ساز تفکیک‌شده Per-Head Muon [۴] و زوال وزن 0.1 آموزش داده شدند.

مشخصات آماری و معماری داده‌محور (Architectural Specifications) جدول ۳ مقایسه مشخصات و پارامترهای پیش‌آموزش مدل **Kimi K3** را در برابر **Kimi K2** نشان می‌دهد. این مشخصات ظرفیت داده‌پذیری شبکه را در سطح ۷۸.۲ تریلیون پارامتر با ۸۹۶ متخصص مسیردهی‌شده تبیین می‌کند.

جدول ۳: مقایسه معماری و مشخصات پیش‌آموزش مدل‌های **Kimi K2** و **Kimi K3**

ویژگی / مؤلفه معماری	Kimi K2	Kimi K3	تغییرات (△)
معماری کلی (Architecture)	MoE	MoE	–
تعداد کل لایه‌ها (#Layers)	۶۱	۹۳	↑ 52%
تعداد کل پارامترها (Total Parameters)	۰۴T.۱	**۷۸T.۲**	↑ 167%
پارامترهای فعال به‌ازای هر توکن (Activated Parameters)	۶B.۳۲	**۲B.۱۰۴**	↑ 220%
ابعاد پنهان لایه‌ها (Hidden Dimension)	۷،۱۶۸	۷،۱۶۸	=
ابعاد لایه فشرده نهفته (Latent MoE Dimension)	–	۳،۵۸۴ (0.5×)	–
ابعاد پنهان به‌ازای هر متخصص (MoE Hidden Dim per Expert)	۲،۰۴۸	۳،۰۷۲	↑ 50%
تعداد کل متخصصان مسیردهی‌شده (Routed Experts)	۳۸۴	**۸۹۶**	↑ 133%
تعداد متخصصان فعال در هر توکن (Active Experts per Token)	۸	**۱۶**	↑ 100%
تعداد متخصصان اشتراکی (Shared Experts)	۱	۲	↑ 100%
تعداد سرهای توجه (Attention Heads)	۶۴	۹۶	↑ 50%
تعداد لایه‌های متراکم (Dense Layers)	۱	۱	=
اندازه دایره واژگان (Vocabulary Size)	۱۶۰K	۱۶۰K	=
طول بافتار پیش‌آموزش (Training Context Length)	۱۲۸K	**۱M**	↑ 8×
مکانیزم توجه (Attention Mechanism)	MLA	**Hybrid KDA–MLA**	–
ترکیب‌بندی لایه‌های توجه	MLA ۶۱	MLA ۲۴ + KDA ۶۹	–
تابع فعال‌ساز (Activation Function)	SwiGLU [۶]	**SiTU-GLU**	–
تعداد لایه‌های پیش‌بینی چندتوکنی (MTP Layers)	layer ۱	layer ۱	=
کل پارامترهای برج بینایی (ViT Total Parameters)	–	**۴۰۱M**	–
تعداد لایه‌های برج بینایی (#ViT Layers)	–	layers ۲۷	–
اندازه پچ ورودی تصویر (Patch Size)	–	۱۴	–
تعداد سرهای توجه برج بینایی (ViT Attention Heads)	–	۱۲	–

- **Magistral**

مقدمه و تبیین ماهیت دادگان مدل مدل‌های خانواده **Magistral** به عنوان نخستین مدل‌های استدلالی شرکت Mistral AI [۱۹] معرفی شده‌اند. برخلاف رویه سنتی پیش‌آموزش خام بر روی تریلیون‌ها توکن وب، این پژوهش مستقیماً از چک‌پوینت‌های پیش‌آموزش‌دیده پایه‌ای شامل Mistral Medium 3 Instruct و Mistral Small 3 Instruct بهره جسته است. تمرکز محوری در این مدل، مهندسی و گردآوری داده‌های استدلالی تاییدپذیر (Verifiable Rewards) برای بهینه‌سازی یادگیری تقویتی (RLVR) و فاز راه‌اندازی نظارت‌شده (SFT Bootstrapping) است:

- **مدل Magistral Medium:** آموزش‌دیده بر بستر Mistral Medium 3 صرفاً از طریق یادگیری تقویتی خالص (Pure RL) بدون بهره‌گیری از هرگونه تقطیر یا ردپای استدلالی اولیه، که منجر به جهش نزدیک به ۵۰ درصدی در بنچمارک AIME-24 گردیده است.
- **مدل Magistral Small:** مدلی ۲۴ میلیارد پارامتری بر پایه Mistral Small 3 که فرایند آموزش آن با تلفیق داده‌های آغازین سرد (Cold-start Data) استخراج‌شده از نسخه Medium و سپس اجرای برخط RL تکمیل شده است.

پالایش چندمرحله‌ای دادگان ریاضی (Math Data Curation) فرایند گردآوری دادگان ریاضی با هدف دست‌یابی به مسائلی با پاسخ‌های تحلیلی، قطعی و ارزیابی‌پذیر توسط مفسرهای نمادین طراحی شده است:

- **پالایش ساختار و فرمت (Format Filtering):** کار با یک مجموعه بزرگ اما نویزدار شامل حدود 700K مسئله آغاز گردید. به منظور امکان‌پذیر ساختن اعتبارسنجی قطعی، تمامی مسائل اثباتی تشریحی و مسائل چندبخشی که ارزیابی آنها با قواعد الگوریتمی دشوار بود، حذف شدند. همچنین پرسش‌های چندگزینه‌ای به مسائل گزاره‌ای و تشریحی بازنویسی شدند تا احتمال حدس تصادفی خنثی شده و استخراج پاسخ در قالب ساختار جبری SymPy درون جعبه `\boxed{}` مقدور گردد. این گام حجم داده‌ها را به 501K کاهش داد.
- **فیلتر دوسطحی دشواری و ناحیه طلایی (Goldilocks Zone Filtering):** برای جلوگیری از افت گرادیان ناشی از نمونه‌های بسیار آسان یا ناممکن، یک غربالگری دومرحله‌ای اعمال شد:

۱. در مرحله نخست، با نمونه‌برداری ۱۶ پاسخ مستقل به ازای هر مسئله توسط مدل Mistral Large 2 [۲۰]، مسائلی با نرخ موفقیت صفر یا صد درصد کنار گذاشته شدند و داده‌های باقی‌مانده صرف آموزش یک چک‌پوینت نمره‌گذار ۲۴ میلیارد پارامتری شدند.

۲. در مرحله دوم، این مدل استدلالی ۲۴ میلیارد پارامتری کل دادگان را مجدداً ارزیابی نمود. مسائلی که اکثریت خروجی‌های مدل در آنها بر پاسخی یکسان اجماع داشتند اما با برچسب واقعیت زمینی (Ground Truth) داده اولیه مغایر بودند، به دلیل خطا در برچسب مرجع حذف گردیدند.

در نتیجه این فیلترینگ سخت‌گیرانه، داده‌های نهایی ریاضی به **۳۸ هزار مسئله فوق‌العاده باکیفیت** محدود شد.

گردآوری و مهندسی دادگان برنامه‌نویسی (Code Data Curation) پیکره کدنویسی رقابتی با هدف سنجش دقیق درستی و کارایی الگوریتم‌ها شکل گرفته است:

- **اعتبارسنجی و تصحیح آزمون‌های واحد (Unit Tests):** مسائل فاقد آزمون یا کدهای مرجع معتبر حذف شدند. راه‌حل‌های مرجع بر روی آزمون‌ها اجرا شده و آزمون‌های ناهماهنگ پالایش شدند. در مواردی که تمامی راه‌حل‌ها در یک تست با شکست مواجه می‌شدند، برچسب تست بر مبنای خروجی متداول راه‌حل‌ها تصحیح شد. در صورت نیاز، تست‌های مصنوعی جدیدی نیز تولید و اضافه گردید.
- **پشتیبانی دوزبانه و استانداردسازی کامپایل:** مسائل به گونه‌ای بازطراحی یا تکثیر شدند که پیاده‌سازی به هر دو زبان Python و C++20 پشتیبانی شود. برای تسریع زمان ارزیابی، هدر پرکاربرد `bits/stdc++.h` پیش‌کامپایل گردید.
- **محیط ارزیابی ایمن (Sandbox Execution):** برای هر ارزیابی گروهی، ۲۰ آزمون تصادفی انتخاب و به تمامی اعضای گروه اختصاص یافت. سقف زمان اجرای هر تست ۴ ثانیه و محدودیت حافظه 300MB تعیین گردید. این خط لوله در نهایت به گردآوری **۳۵ هزار مسئله کدنویسی** انجامید.

مهندسی دادگان چندزبانه و پایداری زبان تفکر (Multilingual Strategy) به منظور مهار پدیده سوئیچ ناخواسته زبان در طول زنجیره تفکر (Language Mixing):

- **ترجمه متوازن دادگان:** دقیقاً 10% از مسائل انگلیسی به زبان‌های فرانسوی، اسپانیایی، ایتالیایی، آلمانی، چینی و روسی برگردانده شد.
- **پاداش پایداری زبان (Language Consistency):** پس از پاک‌سازی علائم لاتک و کدهای دستوری، سه‌گانه پرسش، افکار داخل تگ `<think>` و پاسخ نهایی توسط طبقه‌بند fastText [۲۱] بررسی شده و در صورت یکسانی زبان در هر سه بخش، پاداش افزوده‌ای معادل 0.1 اعمال گردید. این تمهید امکان تولید زنجیره تفکر به زبان بومی کاربر را بدون افت کیفیت استدلال تثبیت کرد.

دادگان راه‌اندازی اولیه و تقطیر برای مدل کوچک (SFT Cold-Start Bootstrapping) در آموزش مدل **Magistral Small**، برای آماده‌سازی اولیه سیاست مدل:

- **استخراج ردپاهای باکیفیت آموزگار:** نمونه پاسخ‌های صحیح از مراحل پیشرفته آموزش RL مدل Magistral Medium استخراج و ردهای اولیه با تفکر کوتاه فیلتر شدند.
- **غنی‌سازی با داده‌های متن‌باز:** پرومپت‌های متنوع از مجموعه‌های OpenThoughts [۲۲] و OpenR1 [۲۳] گزینش شده و با پاسخ‌های تفکری مدل Medium غنی‌سازی شدند.
- **تزریق داده‌های عمومی دستورپذیری:** برای پیشگیری از فراموشی فاجعه‌بار مهارت‌های عمومی، 10% از حجم داده‌های SFT به دستورات متداول غیراستدلالی اختصاص یافت و مدل به مدت ۴ دوره بر روی این ترکیب آموزش دید.

صورت‌بندی بهینه‌سازی و سیگنال‌های پاداش حاکم بر دادگان فرایند بهینه‌سازی سیاست مدل π_θ بر روی دادگان پردازش‌شده توسط الگوریتم GRPO [۲۴] و با اعمال اصلاحات تثبیت‌کننده به فرم زیر اجرا می‌شود:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim\mathcal{P}(Q),\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}}\left[\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\min\left(r_{i,t}(\theta)\hat{A}_{i,t}^{\text{norm}}, \text{clip}(r_{i,t}(\theta), 1-\varepsilon_{\text{low}}, 1+\varepsilon_{\text{high}})\hat{A}_{i,t}^{\text{norm}}\right)\right] \tag{۱۱}$$

که در آن $r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|q,o_{i,<t})}$، نسبت مزیت استانداردشده در سطح مینی‌بچ به صورت $\hat{A}_{i,t}^{\text{norm}} = \frac{\hat{A}_i - \hat{A}_{\text{mean}}}{\hat{A}_{\text{std}}}$ تعریف شده و شرط تفکیک گروه‌های همگن $\exists m < n, r_m \neq r_n$ اعمال می‌شود.

همچنین جهت پیشگیری از رشد مهارگسیخته طول پاسخ و سرریز حافظه، جریمه طول نرم بر داده‌ها حاکم گردید:

$$R_{\text{length}}(y) = \begin{cases} 0, & |y| \leq l_{\text{max}} - l_{\text{cache}} \\ -0.1 \cdot \frac{|y| - l_{\text{max}} + l_{\text{cache}}}{l_{\text{cache}}}, & l_{\text{max}} - l_{\text{cache}} < |y| \leq l_{\text{max}} \\ -0.1, & l_{\text{max}} < |y| \end{cases} \tag{۱۲}$$

تحلیل‌های آماری، پویایی وزن‌ها و ارزیابی جامع دادگان داده‌های آموزشی و اثرات آنها بر روی شاخص‌های کمی در جداول زیر خلاصه شده است. جدول ۴ فرآیند فشرده‌سازی و غربالگری دادگان ریاضی را نشان می‌دهد که در آن تنها ۴٪.۵ از داده‌های اولیه موفق به عبور از فیلترهای صحت و دشواری شدند. جدول ۵ نیز نشان‌دهنده ارتقای چشمگیر عملکرد استدلالی مدل در حوزه‌های گوناگون تحت آموزش بر روی این دادگان گزینش‌شده است.

یافته‌های انتقال بدون داده و راهبردهای ناموفق (Negative Results) تحلیل پویایی آموزش مدل حاوی دو بینش اساسی در حوزه مهندسی دادگان است:

- **انتقال چندوجهی بدون دادگان تصویر (Multimodal Free Lunch):** علیرغم اینکه آموزش RL کاملاً بر بستر داده‌های متنی محض صورت پذیرفت، مدل قابلیت‌های انکودر بصری پیش‌آموزش‌دیده خود را حفظ کرده و حتی در بنچمارک‌های چندوجهی ارتقا یافت؛ به گونه‌ای که امتیاز MMMU از 65.0% به 70.0% و در MMMU-Pro (Vision) با رشدی چشمگیر از 39.7% به 52.1% رسید.

جدول ۴: مراحل متوالی پالایش کمی دادگان ریاضی در مدل‌های **Magistral**

شاخص	داده اولیه خام (Initial)	پالایش فرمت و ساختار	پالایش دشواری ناحیه طلایی
تعداد نمونه‌ها	699K	501K	38K
نرخ بقا نسبت به مرحله قبل	-	71.67%	7.58%
نرخ تجمعی بقا از کل	100%	71.67%	**۴۴%.۵**

جدول ۵: ارزیابی عملکرد مدل‌های **Magistral** در مقایسه با مبناهای استدلالی تراز اول جهان

بنچمارک (متریک)	Mistral Med. 3	Magistral Med.	DeepSeek-V3	DeepSeek-R1-Zero	DeepSeek-R1
استفاده از SFT قبل از RL	-	**خیر**	-	خیر	بله
AIME ’24 (Pass@1 / Maj@64)	۴.۴۳ / ۸.۲۶	**۰.۹۰ / ۶.۷۳**	۲.۳۹	- / ۰.۷۱	- / ۸.۷۹
AIME ’25 (Pass@1 / Maj@64)	۰.۳۰ / ۲.۲۱	**۳.۸۳ / ۹.۶۴**	۸.۲۸	-	- / ۰.۷۰
MATH-500	۰.۹۱	**۳.۹۴**	۲.۹۰	۹.۹۵	۳.۹۷
GPQA Diamond	۶.۵۹	**۸.۷۰**	۱.۵۹	۳.۷۳	۵.۷۱
LiveCodeBench v5	۱.۲۹	**۴.۵۹**	۲.۳۶	۰.۵۰	۹.۶۵
Aider Polyglot	۹.۲۸	**۱.۴۷**	۶.۴۹	-	۳.۵۳
Humanity’s Last Exam (Text)	۴.۴	**۰.۹**	-	-	۶.۸

- **شکست پاداش جزئی در دادگان کد:** فرضیه اعطای پاداش نسبی بر پایه درصد قبولی آزمون‌های واحد به جای پاداش دودویی (0 یا 1)، به افت ۲ درصدی دقت در LiveCodeBench انجامید، زیرا پاداش جزئی منجر به تقویت الگوهای غیربهینه و ایجاد سیگنال گمراه‌کننده در کدهای دارای باگ می‌شد.

- **MiniMax-M2**

مقدمه و ساختار کلی دادگان پیش‌آموزش فاز پیش‌آموزش (Pre-training) در مدل زبانی **MiniMax-M2** با تکیه بر اصل راهنمای «فعال‌سازی حداقلی برای آزادسازی بیشترین هوش دنیای واقعی» (Mini Activations Unleashing Max Real-World Intelligence) طراحی شده است. معماری این مدل یک ترنسفورمر فقط دکودر با ۹.۲۲۹ میلیارد پارامتر کل است که در هر گام محاسباتی تنها ۸.۹ میلیارد پارامتر فعال می‌گردد (۲۵۶ متخصص ریزدانه با مسیریابی سیگموئید و فعال‌سازی هم‌زمان ۸ متخصص). فرایند پیش‌آموزش پایه بر روی مجموعه‌ای بالغ بر **۲.۲۹ تریلیون توکن** (29.2T tokens) با سقف بافت بومی ۱۹۲ هزار توکن (192K native context window) صورت پذیرفته است [۲۵]. این بودجه کل توکن‌ها به صورت دومرحله‌ای بهینه‌سازی شده است:

$$|\mathcal{D}_{\text{pretrain}}| = |\mathcal{D}_{\text{constant}}| + |\mathcal{D}_{\text{decay}}| = 19.9\text{T} + 9.3\text{T} = 29.2\text{T} \quad (۱۳)$$

- **فاز نرخ یادگیری ثابت (Constant Phase):** شامل ۹.۱۹ تریلیون توکن (معادل ۱۵%.۶۸ از کل داده‌ها) که تمرکز آن بر تثبیت بازنمایی‌های دانش عمومی، مبانی استدلال و یادگیری ساختار زبان در طول بافت‌های اولیه (8K) است.
- **فاز زوال و بسط زمینه (Decay Phase):** شامل ۳.۹ تریلیون توکن (معادل ۸۵%.۳۱ از کل داده‌ها) که ضمن کاهش نرخ یادگیری، با تزریق داده‌های سنتزشده بافت‌بلند، تعمیم مدل تا سقف 192K توکن را محقق می‌سازد.

ترکیب‌بندی و پالایش پیکره داده‌ها (Data Construction & Filtering) پیکره متنی پیش‌آموزش از منابع متعددی شامل متون پالایش‌شده وب، مقالات دانشگاهی، کتاب‌ها، کدهای برنامه‌نویسی و دادگان ساختاریافته پرسش و پاسخ گردآوری شده است. با توجه به هدف‌گذاری مدل بر روی رفتارهای عاملی (Agentic Workflows)، تدابیر کیفی زیر اعمال گردید:

- **بیش‌نمونه‌برداری راهبردی (Domain Upsampling):** داده‌های مربوط به کد (Code)، ریاضیات (Mathematics) و علوم پایه، فناوری، مهندسی و ریاضیات (STEM) به صورت چشمگیری فراتر از نرخ توزیع طبیعی خود در وب بیش‌نمونه‌برداری شدند تا توانایی استنتاج منطقی ارتقا یابد.

- **پالایش دوبخشی کیفیت داده‌ها (Model-Based Quality Filtering):** اسناد خام ورودی از دو فیلتر ترکیبی عبور می‌کنند: ۱) مدل‌های امتیازدهی پاداش متنی (Model-Based Reward Scoring) جهت سنجش چگالی اطلاعاتی و سبک نگارش؛ ۲) طبقه‌بندهای کمکی چندبعدی (Aux-iliary Classifiers) جهت اعتبارسنجی تخصصی و موضوعی اسناد.
- **استراتژی نمونه‌برداری متعادل (Balanced Sampling Strategy):** به منظور پرهیز از افت تنوع موضوعی، داده‌های دارای رتبه کیفی بالا وزن‌دهی مثبت شده (Upweighting)، در حالی که پوشش جامع زبان‌شناختی و دانش حاشیه‌ای جهان به دقت محافظت می‌گردد.

پیش‌بینی چندتوکنی و اهداف آموزشی (Multi-Token Prediction & Objectives) به منظور غنی‌سازی سیگنال‌های نظارتی در طول مصرف توکن‌های پیش‌آموزش و تسریع استنتاج از طریق کدگشایی فرضی (Speculative Decoding) [۲۶]، ماژول پیش‌بینی چندتوکنی (MTP) [۲۷، ۱] در طول آموزش فعال گردیده است:

$$\mathcal{L}_{\text{total}}(\theta) = \mathcal{L}_{\text{NTP}}(\theta) + \lambda_{\text{MTP}} \sum_{k=1}^{K} \mathcal{L}_{\text{MTP}}^{(k)}(\theta) \tag{۱۴}$$

- **زمان‌بندی وزن زیان (MTP Loss Weight Annealing):** در فاز پیش‌آموزش اولیه با یک ماژول ($K = 1$)، ضریب تلفات ابتدا روی $\lambda_{\text{MTP}} = 0.3$ تنظیم شده و در طول فاز زوال تا سقف 0.1 کاهش می‌یابد.
- **توسعه ماژول با کپی وزن‌ها (Expansion via Weight Copying):** در ادامه پیش‌آموزش، ماژول MTP با کپی وزن‌های مدل اصلی به سه ماژول ($K = 3$) ارتقا می‌یابد. در گام نخست، مدل اصلی فریز شده و فقط ماژول‌های MTP تا تثبیت خطا آموزش می‌بینند و سپس آموزش مشترک کل شبکه از سر گرفته می‌شود.

برنامه درسی طول بافت و داده‌های بلندکانتکست (Context Curriculum & Long-Context Data) فرایند تعمیم مدل به بافت فوق‌طولانی 192K توکنی در طول فاز زوال (9.3T توکن) با برنامه آموزشی تدریجی زیر صورت پذیرفته است:

$$8\text{K} \longrightarrow 32\text{K} \longrightarrow 192\text{K} \tag{۱۵}$$

به منظور پوشش بهینه پنجره‌های زمینه گسترده بدون اتلاف محاسباتی، سه رویکرد مهندسی داده پیاده‌سازی شده است:

- **الحاق کدهای پروژه‌ای باکیفیت (Code Concatenation):** تجمیع فایل‌های متنی وابسته در مخازن کد به نحوی که روابط ارجاعی و ساختار پروژه حفظ گردد.
- **اسناد بلند ذاتاً ساختاریافته (Naturally Long-Form PDFs):** انتخاب گزارش‌های تحلیلی، متون قانونی و کتب با پیوستگی استدلالی طبیعی در طول توالی‌های بالا.
- **بسته‌بندی اسناد هم‌موضوع (Thematically Related Document Packing):** ادغام هوشمند اسناد دارای ارتباط معنایی در یک توالی واحد به منظور حذف کامل توکن‌های پرکننده (Padding) و تمرین استدلال بین‌سندی.

ارزیابی‌های تجربی و مطالعات ابطال در مقیاس پیش‌آموزش (Pre-Training Evaluations & Ablations) کیفیت پیکره پیش‌آموزش و انتخاب‌های معماری مدل پایه **MiniMax-M2** طی دو ارزیابی جامع سنجیده شده است:

۱. مطالعه ابطال روی بودجه ۵۰۰ میلیارد توکنی: در جدول ۶، کارایی یادگیری از داده‌های پیش‌آموزش تحت سه پیکربندی مجزا بر روی ۵۰۰ میلیارد توکن سنجیده شده است. نتایج حاکی از آن است که ترکیب متخصصان ریزدانه (Fine-Grained) [۲۸] و ماژول MTP [۲۷] با ثابت نگه داشتن حجم داده‌ها

جدول ۶: مطالعه ابطال طراحی متخصصان ریزدانه و ماژول MTP بر روی ۵۰۰ میلیارد توکن پیش‌آموزش [۲۵]

مشخصات / بنچمارک	تنظیمات (Shots)	پایه (Baseline)	همراه با MTP	همراه با Fine-Grained
پارامترهای فعال	–	۲B	۲B	۲B
کل پارامترها	–	۸B.۱۷	۸B.۱۷	۸B.۱۷
بودجه توکن‌های آموزش	–	۵۰۰B	۵۰۰B	۵۰۰B
تعداد کل متخصصان (num_experts)	–	۳۲	۳۲	۱۲۸
تعداد متخصصان فعال (topk)	–	۲	۲	۸
[۲۹] MATH	۴-shot	۶.۱۹	۳.۲۱	**۱.۲۴**
[۳۰] MMLU	۵-shot	۸.۳۹	۷.۳۹	**۲.۴۰**
[۳۱] ARC-Challenge	۲۵-shot	۴.۲۷	۵.۲۷	**۸.۲۷**
[۳۲] KorBench	۳-shot	۱.۱۴	**۰.۱۵**	۸.۱۴
[۳۳] HumanEval	۰-shot	۷.۲۹	۱.۳۰	**۵.۳۲**

و پارامترهای فعال (2B)، بازدهی استخراج دانش از داده‌ها را در بنچمارک استدلال ریاضی MATH از 19.6 به 24.1 (رشد نسبی ۲۲٪) و در کدنویسی HumanEval از 29.7 به 32.5 ارتقا داده است.

۲. مقایسه توجه کامل در برابر توجه پنجره‌ای لغزان (Full Attention vs. Hybrid SWA): تیم **Min-iMax** برای بهینه‌سازی پردازش داده‌های بلندکانتکست، مکانیزم توجه پنجره‌ای لغزان ترکیبی (Hybrid SWA) [۳۴] را در مقیاس معماری **M2** در برابر توجه کامل چندسری (Full Attention) [۳۵، ۳۶] مورد آزمون قرار داد. مطابق جدول ۷، علیرغم هم‌سطح بودن عملکرد در کانتکست‌های کوتاه (32K)، مکانیزم SWA در بنچمارک‌های بافت‌بلند نظیر RULER 128K CWE با افت ۱۸ نمره‌ای ($90.0 \rightarrow 72.0$) و در ترجمه کانتکست‌محور MTOB با افت شدید مواجه شد. این تحلیل آماری ثابت کرد که بهره‌برداری کامل از داده‌های بلندکانتکست در پیش‌آموزش مستلزم حفظ ساختار توجه کامل است.

جدول ۷: ارزیابی پیش‌آموزش در مقیاس معماری **MiniMax-M2**: توجه کامل در برابر توجه پنجره‌ای لغزان (Hybrid SWA) [۲۵]

دسته آزمون	بنچمارک	پایه: توجه کامل (Baseline)	همراه با SWA
دانش عمومی و استدلال	[۳۷] HELMET ICL	**۸.۷۵**	۷.۷۲
	[۳۰] MMLU	۵.۸۵	**۶.۸۵**
	[۲۹] MATH	**۳.۶۰**	**۳.۶۰**
بازیابی در بافت طولانی	[۳۸] RULER 128K CWE	**۰.۹۰**	۰.۷۲
	[۳۸] RULER 128K MQ	**۰.۹۹**	۰.۹۳
	[۳۸] RULER 32K CWE	**۰.۹۹**	**۰.۹۹**
	[۳۸] RULER 32K MQ	**۰.۹۹**	**۰.۹۹**
ترجمه درون‌بافتی	[۳۹] MTOB K-e Bleurt	**۰.۶۰**	۰.۴۵
	[۳۹] MTOB e-k ChrF	**۸.۴۴**	۲.۲۷

- **MobileLLM-Pro**

مقدمه و ساختار چندمرحله‌ای دادگان پیش‌آموزش مدل **MobileLLM-Pro** [۴۰] یک مدل زبانی بنیادین ۱ میلیارد پارامتری (دقیقاً ۰۸۴.۱ میلیارد پارامتر) است که با هدف استقرار و استنتاج کارآمد روی دستگاه‌های لبه‌ای (On-Device) طراحی شده است. از آنجا که مدل‌های با مقیاس زیر ۳ میلیارد پارامتر حساسیت بسیار بالایی نسبت به نویز دادگان و پدیده جابجایی توزیع (Distribution Shift) در طول مراحل گوناگون آموزش دارند، فرایند پیش‌آموزش داده‌ها در این مدل بر پایه یک خط‌لوله ۴ فازی مهندسی شده است:

- **فاز ۱ (اکتساب زبان پایه - Language Acquisition):** تمرکز بر یادگیری عمومی زبان با بودجه آموزشی **۴.۱ تریلیون توکن** (1.4T tokens) و طول بافت 2K توکن.
- **فاز ۲ (بسط بافت متنی - Context Expansion):** توسعه پنجره بافت به **۱۲۸،۰۰۰ توکن** (128K context) روی **۲۰ میلیارد توکن** بدون تغییر در توزیع متنی فاز ۱.

- **فاز ۳ (بازپخت و ادغام متخصصان - Specialist Model Merging):** بازپخت مدل با خوشه‌بندی دادگان پیش‌آموزش روی حوزه‌های هدف با بودجه **۶۰ میلیون توکن** برای هر شاخه متخصص.
- **فاز ۴ (آموزش آگاه از کوانتایش - Quantization-Aware Training):** فشرده‌سازی و پایدارسازی وزن‌های مدل تا سطح ۴ بیتی بر روی **۸۰ میلیارد توکن** (حدود ۵٪ از بودجه کل آموزش).

در تمامی فازهای پیش‌آموزش، به‌جای استفاده از اهداف کلاسیک تک‌مقداری (One-hot targets) و تابع زیان انتروپی متقاطع مرسوم، از تقطیر دانش مبتنی بر لاجیت (Logit-based Knowledge Distillation) [۴۱، ۴۲] از مدل معلم قدرتمند Llama 4-Scout با واژگان کامل ۲۰۲،۰۴۸ توکنی استفاده شده است. هدف بهینه‌سازی از طریق حداقل‌سازی واگرایی پیشرو کولبک-لایبلر (Forward KL-Divergence) تعریف می‌شود:

$$\mathcal{L}_{\mathsf{KD}}(\theta) = \mathcal{D}_{\mathsf{KL}}\big(P_{\mathsf{Teacher}}(y \mid x) \,\|\, P_{\mathsf{Student}}(y \mid x;\theta)\big) \tag{۱۶}$$

شبیه‌سازی آفلاین و گزینش بهینه داده‌ها (SDM Data-Mix Simulations) به منظور پرهیز از تصمیم‌گیری‌های شهودی یا وزن‌دهی یکنواخت، ترکیب داده‌های فاز ۱ با استفاده از چهارچوب شبیه‌ساز مقیاس‌پذیر داده (Scalable Data Mixer - SDM) [۴۳، ۴۴] استخراج شده است. در این شیوه، اثرگذاری هر نمونه متنی x بر دامنه‌های گوناگون داده‌ای D (شامل زبان عمومی، دانش، استدلال، کد و ریاضیات) از طریق مدل‌های زبانی آماری سبک‌وزن (SLM) سنجیده می‌شود:

$$\mathsf{Inf}(x, D) = \mathcal{L}_D(\mathsf{SLM}) - \mathcal{L}_D(\mathsf{SLM} \leftarrow x) \tag{۱۷}$$

که در آن $\mathcal{L}_D$ زیان انتروپی متقاطع روی دامنه D را قبل و بعد از آموزش یک‌گامه با نمونه x نشان می‌دهد. این امتیازها در قالب یک بردار تأثیر $\mathbf{\Delta}_x = [\mathsf{Inf}(x, D_1), \ldots, \mathsf{Inf}(x, D_k)]^T$ تجمیع شده و توسط یک رگرسور عصبی سبک‌وزن، مطلوبیت پایین‌دستی (Downstream Utility) $\hat{y}_D$ پیش‌بینی می‌شود. در نهایت، با یک بار پیمایش آفلاین کل پیکره در طول ۱ میلیون گام، وزن‌های ایستای نمونه‌برداری w_i متناسب با مطلوبیت انتظاری $\mathbb{E}_{x\in\mathcal{D}_i}[\hat{y}]$ به دست می‌آیند به‌طوری که $\sum_i w_i = 1.0$.

آمار تفکیکی دادگان پالایش‌شده برای پیش‌آموزش فاز ۱ و ۲ در جدول زیر خلاصه شده است:

جدول ۸: مشخصات آماری دادگان فاز ۱ و فاز ۲ پیش‌آموزش مدل **MobileLLM-Pro**

مجموعه داده (Dataset)	مرجع علمی	سطرها (میلیون)	حجم توکن (میلیارد)	وزن در آمیزه (%)
Fineweb-EDU (2024)	[۴۵] al. et Allal	۱.۱۲۷۹	۰.۱۳۰۰	**۷۵%.۸۹**
Starcoder (2023)	[۴۶] al. et Li	۶.۲۰۶	۸.۲۶۳	**۶۶%.۴**
Open Web Math (2023)	[۴۷] al. et Paster	۱.۶	۶.۱۲	**۹۲%.۱**
Arxiv (2023)	-	۵.۱	۰.۲۸	**۳۵%.۱**
Wikipedia	-	۲.۷	۷.۳	**۰۲%.۱**
Stack Exchange (2023)	[۴۸] Computer Together	۲.۲۹	۶.۱۹	**۰۲%.۱**
Algebraic Stack (2023)	[۴۹] al. et Azerbayev	۴.۳	۶.۱۲	**۲۴%.۰**
مجموع کل (Total)	-	**۰.۱۵۰۰**	**۳.۱۶۴۰**	**۰۰%.۱۰۰**

این توزیع آماری نشان می‌دهد که مدل بیشترین سهم خود (۷۵٪.۸۹) را از متون آموزشی و منسجم وب دریافت کرده و تخصیص متوازن ۸۲٪.۶ از کل آمیزه به داده‌های استدلالی، برنامه‌نویسی و ریاضیات، ظرفیت محاسباتی مدل ۱ میلیارد پارامتری را بدون اشباع یا نویزپذیری پایدار ساخته است. این مرحله با اندازه دسته موثر ۲ میلیون توکن در ۶۴۰،۰۰۰ گام به‌روزرسانی و با نرخ یادگیری اولیه 4×10^{-4} آموزش داده شده است.

بسط بافت متن بدون جابجایی توزیع (Implicit Positional Distillation) یکی از نوآوری‌های بنیادین در سازمان‌دهی داده‌های پیش‌آموزش **MobileLLM-Pro**، حذف نیاز به پیکره‌های متنی بلند مرسوم در فاز دوم است. تزریق متون فوق‌طولانی معمولاً تعادل زبانی مدل‌های فشرده را برهم می‌زند و به افت

یادگیری منجر می‌شود. در این پژوهش، از تکنیک «تقطیر موقعیتی پنهان» بر بستر تعبیه موقعیتی چرخشی (RoPE) استفاده شده است:

$$\begin{bmatrix} x'_i \\ x'_{i+1} \end{bmatrix} = \begin{bmatrix} \cos(\theta_i p) & -\sin(\theta_i p) \\ \sin(\theta_i p) & \cos(\theta_i p) \end{bmatrix} \begin{bmatrix} x_i \\ x_{i+1} \end{bmatrix}, \quad \theta_i = 10000^{-\frac{2i}{d}} \tag{۱۸}$$

با استفاده از اتصال اسناد کوتاه قبلی و پوشش علیتی بلوکی (Block Causal Masking)، مدل دانشجو توزیع لاجیت معلم را دریافت می‌کند. از آنجا که مدل معلم قبلاً روی توالی‌های بلند آموزش دیده، لاجیت‌های آن رفتار چرخش زاویه‌ای کامل (تا سقف ۳۶۰ درجه معادل بافت ۱۲۸K) را به صورت پنهان بازتاب می‌دهند. در نتیجه، بدون مواجهه مستقیم مدل دانشجو با متون طولانی اختصاصی و تنها با آموزش روی **۲۰ میلیارد توکن مضاعف** از همان توزیع داده‌ای فاز اول (طی ۱۰۰،۰۰۰ گام با بیشینه نرخ یادگیری 4×10^{-5})، مدل به پنجره بافت ۱۲۸K دست پیدا می‌کند.

خوشه‌بندی داده‌ها و ادغام متخصصان (Specialist Model Merging) در فاز سوم، داده‌های بدون ساختار پیش‌آموزش خوشه‌بندی شده و به زیرمجموعه‌های متمرکز بر حوزه‌های خاص تفکیک می‌شوند. به‌جای آموزش هم‌زمان که موجب رقابت مخرب بین قابلیت‌ها (نظیر حفظ حقایق در برابر توانایی استدلال) می‌شود، n مدل متخصص موازی $(b_1, \ldots, b_n)$ هر کدام به‌طور مستقل روی دامنه متنی اختصاصی خود $\mathcal{D}_i$ به میزان **۶۰ میلیون توکن** طی **۵۰۰ گام** آموزش می‌بینند. در نهایت، پارامترهای متخصصان از طریق ادغام غیریکنواخت وزنی تلفیق می‌شوند:

$$M = \frac{1}{n} \sum_{b=1}^{n} \theta_b \cdot w_b, \quad \sum_{b=1}^{n} w_b = 1.0 \tag{۱۹}$$

این رویکرد داده‌محور، مدل پایه **MobileLLM-Pro-base** را حاصل می‌آورد که در آن ویژگی‌های مجزای حوزه‌ای به‌شکل هم‌افزا پیوند خورده‌اند.

بودجه داده‌ها در فاز کوانتایش (QAT Phase) در مرحله چهارم، به منظور انطباق نهایی وزن‌ها و فعال‌سازها برای استقرار روی پردازنده‌های کم‌مصرف CPU و شتاب‌دهنده‌های NPU (مانند Apple ANE و Qualcomm HTP)، مدل با بودجه متنی معادل **۸۰ میلیارد توکن** با روش تقطیر از خود (Self-Distillation) از نسخه اعشاری فاز ۳، تحت یادگیری بازه کوانتایش قرار می‌گیرد.

ارزیابی‌های تجربی و تحلیل آماری اثرگذاری داده‌ها جدول‌های زیر اثرات تجربی راهبردهای انتخاب، پالایش و عدم جابجایی توزیع داده‌ها را بر عملکرد نهایی مدل اثبات می‌کنند:

جدول ۹: مقایسه عملکرد آمیزه داده بهینه‌شده با SDM در برابر آمیزه یکنواخت (Uniform)

مرحله آموزش	بودجه توکن	منبع آمیزه داده	میانگین عملکرد (%) (Avg. Score)
پیش‌آموزش (Pre-training)	۴T.۱	یکنواخت (Uniform) **SDM (پیشنهادی)**	۷۰%.۳۸ **۳۱%.۴۹**
آموزش دستورپذیری (IFT)	۸۰B	یکنواخت (Uniform) **SDM (پیشنهادی)**	۹۴%.۱۷ **۲۳%.۴۵**

بر اساس نتایج جدول فوق، بهینه‌سازی داده‌ها از طریق SDM موجب ارتقای مطلق **۶۱%.۱۰** در مرحله پیش‌آموزش نسبت به توزیع یکنواخت شده است.

همچنین جدول زیر، مزیت ثابت نگه داشتن توزیع داده در فاز ۲ را در مقایسه با تزریق متون طولانی متداول اثبات می‌کند:

همان‌طور که مشاهده می‌شود، استفاده از دادگان متنی طولانی اختصاصی موجب افت ۸۸%.۵ در عملکرد عمومی مدل گردیده، در حالی که روش تقطیر پنهان با حفظ توزیع داده‌های فاز ۱، نمره بازیابی NIH را به ۷۸%.۹۹ رسانده و از افت عملکرد عمومی ممانعت نموده است.

جدول ۱۰: ارزیابی راهبردهای دادگان فاز ۲ در آزمون سوزن در انبار کاه (NIH) و میانگین تسک‌های عمومی

پیکربندی داده‌ای فاز ۲	شاخص NIH (%)	میانگین عملکرد عمومی (بدون NIH)
فاز ۱ (طول بافت کوتاه 2K)	۷۰%.۶	۷۴%.۵۳
فاز ۲ با متون طولانی متعارف	۲۲%.۸۰	۸۶%.۴۷ (افت ۸۸٪.۵)
فاز ۲ با تقطیر پنهان (توزیع داده‌های فاز ۱)	**۷۸%.۹۹**	**۵۷%.۵۳** (حفظ ثبات کیفی)

در نهایت، جدول زیر نشان‌دهنده تأثیر خوشه‌بندی داده‌های پیش‌آموزش روی دامنه‌های تخصصی مستقل و سپس ادغام وزنی آن‌ها است:

جدول ۱۱: تأثیر خوشه‌بندی دامنه‌ها روی متخصصان مجزا و هم‌افزایی حاصل از ادغام وزنی نهایی (Weight Avg)

بنچمارک (متریک)	پیش از بازپخت	متخصص ۱	متخصص ۲	متخصص ۳	متخصص ۴	ادغام نهایی (Weight Avg.)
HellaSwag (acc_char)	۷۷.۶۵	۶۹.۶۶	۸۷.۶۶	۳۱.۶۶	۰۴.۶۷	**۵۸.۶۷**
BoolQ (acc_char)	۲۸.۷۱	۵۱.۷۶	۹۴.۷۶	۶۴.۷۳	۴۷.۷۸	**۸۲.۷۶**
PIQA (acc_char)	۵۷.۷۵	۱۴.۷۵	۵۰.۷۶	۴۶.۷۵	۱۷.۷۶	**۵۵.۷۶**
SIQA (acc_char)	۲۹.۴۷	۹۰.۴۷	۹۷.۵۰	۸۸.۴۶	۳۱.۵۰	**۲۳.۵۱**
TriviaQA (em)	۶۱.۳۶	۷۸.۳۶	۰۷.۴۰	۴۸.۳۷	۹۱.۳۹	**۱۸.۴۰**
Natural Questions (em)	۰۲.۱۲	۲۲.۱۲	۸۴.۱۵	۸۳.۱۱	۰۱.۱۶	**۲۹.۱۶**
ARC-Challenge (acc_char)	۶۴.۵۰	۰۵.۵۳	۱۹.۵۲	۳۳.۵۱	۰۵.۵۳	**۱۶.۵۴**
ARC-Easy (acc_char)	۳۷.۷۱	۲۲.۷۵	۴۵.۷۶	۶۶.۷۳	۷۴.۷۶	**۹۶.۷۶**
WinoGrande (acc_char)	۳۰.۶۳	۰۶.۶۳	۱۴.۶۳	۹۰.۶۲	۰۶.۶۳	**۵۴.۶۳**
OBQA (acc_char)	۸۰.۴۱	۴۰.۴۱	۴۰.۴۳	۸۰.۴۱	۴۰.۴۳	**۰۰.۴۴**
NIH (em)	۰۰.۱۰۰	۰۰.۱۰۰	۰۰.۱۰۰	۰۰.۱۰۰	۰۰.۱۰۰	**۰۰.۱۰۰**
میانگین (بدون NIH)	**۵۷.۵۳**	**۸۰.۵۴**	**۲۴.۵۶**	**۱۳.۵۴**	**۴۲.۵۶**	**۷۳.۵۶**

این آمار مؤید آن است که آموزش شاخه‌ای روی داده‌های تفکیک‌شده و ادغام متعاقب پارامترها، امتیاز مدل را از ۵۷.۵۳ به **۷۳.۵۶** ارتقا داده که حتی از بهترین متخصص منفرد (۴۲.۵۶) نیز فراتر رفته و هم‌افزایی بازنمایی دادگان را ثابت می‌کند.

- **Nemotron 3**

مقدمه و ساختار مقیاس دادگان پیش‌آموزش خانواده مدل‌های زبانی **Nemotron 3** (شامل نسخه‌های Nano، Super و Ultra) با تمرکز بر تعاملات کارگزارمحور (Agentic AI)، پنجره زمینه فوق‌طولانی تا سقف ۱ میلیون توکن (1M Context Length) و بیشینه‌سازی توان عملیاتی استنتاج طراحی شده‌اند. زیرساخت داده‌های پیش‌آموزش (Pre-training) این خانواده بر اساس مقیاس‌های محاسباتی و داده‌ای زیر سازمان‌دهی شده است:

- **مقیاس پیش‌آموزش پایدار با فرمت عددی NVFP4:** معماری پیوندی مانبا-متخصصان (Mamba-MoE) این مدل با پایداری کامل تا سقف **۲۵ تریلیون توکن** (25T tokens) با فرمت عددی NVFP4 آموزش داده شده است.
- **سبد داده‌های بازنشر عمومی:** انویدیا متعهد به انتشار وزن‌ها، نرم‌افزارهای آموزش و مجموعه‌ای بالغ بر **۱۰ تریلیون توکن** (>10T tokens) از دادگان پیش‌آموزش تحت مجوزهای باز شده است.
- **دادگان ارزیابی و آزمایش‌های مقایسه‌ای (Ablation Baselines):** برای سنجش دقیق نوآوری‌هایی نظیر معماری LatentMoE و پیش‌بینی چندتوکنی (MTP)، مدل‌های مقایسه‌ای با ۸ میلیارد پارامتر فعال و ۷۳ میلیارد پارامتر کل بر روی یک پیکره همگن به حجم **۱ تریلیون توکن** (1T tokens) پیش‌آموزش دیده‌اند.

تابع هدف پایه در فاز پیش‌آموزش، کمینه‌سازی میانگین منفی لگاریتم درست‌نمایی (NLL) برای

پیش‌بینی توکن‌های آتی است:

$$\mathcal{L}_{\mathsf{LM}}(\theta) = -\frac{1}{N}\sum_{t=1}^{N}\log P_\theta(x_t \mid x_1, x_2, \ldots, x_{t-1}) \tag{۲۰}$$

ترکیب‌بندی، گونه‌شناسی و دادگان ساختگی (Data Typology & Synthetic Data) با توجه به ماهیت استدلالی و لزوم پردازش زمینه‌های چند صد هزار توکنی، داده‌های پیش‌آموزش شامل شاخه‌های متمایز زیر هستند:

- **کدهای منبع در سطح مخزن (Repository-Level Code):** بهره‌گیری از توالی‌های کامل پروژه‌های نرم‌افزاری با طول فراتر از **۱ میلیون توکن** جهت ایجاد فهم عمیق از ساختارهای چندفایلی، وابستگی‌های متقابل و مهندسی نرم‌افزار.
- **پیکره‌های طولانی گفت‌وگو و بازیابی (RAG & Dialogue):** آماده‌سازی و چینش اسناد تخصصی چندگانه و تاریخچه‌های طولانی تعامل کارگزاران.
- **تزریق داده‌های مصنوعی هدفمند (Targeted Synthetic Data):** برای فاز پیش‌آموزش مداوم (CPT)، داده‌های مصنوعی سنتزشده با تمرکز بر چهار محور کلیدی طراحی شدند:
  ۱. بازیابی اطلاعات در فواصل دوربرد (Long-Range Retrieval)
  ۲. استنتاج و استدلال چندمرحله‌ای (Multi-Hop Reasoning)
  ۳. تجمیع و خلاصه‌سازی داده‌ها از چندین سند مستقل (Multi-Document Aggregation)
  ۴. تفکر ترکیبی پیوسته در بافت‌های گسترده

راهبرد پیش‌آموزش مداوم و مهندسی طول بافت (Continued Pre-training - CPT) برای گسترش پنجره بافت به ۱ میلیون توکن در مدل **Nemotron 3 Nano**، پروتکل داده‌ای ویژه‌ای به کار گرفته شد:

- **طول توالی در پیش‌آموزش مداوم:** فاز CPT مستقیماً روی توالی‌هایی به طول 512**K توکن** پیاده‌سازی شده است.
- **حذف برنامه‌ریزی مرحله‌ای طول توالی (No Staged Schedule):** برخلاف روال‌های مرسوم که طول توالی به تدریج از 8K به مقادیر بالاتر ارتقا می‌یابد، در معماری پیوندی این مدل داده‌های بافت 512K بدون نیاز به افزایش گام‌به‌گام به شبکه تزریق شدند.
- **حذف نشانه‌گذاری موقعیتی دوار (RoPE Elimination):** با توجه به اینکه لایه‌های Mamba-2 به صورت ذاتی و ساختاری اطلاعات موقعیت نسبی توکن‌ها را درون متغیرهای حالت پنهان کدگذاری می‌کنند [۵۰]، نیازی به تعبیه RoPE در لایه‌های توجه وجود نداشته و در نتیجه پدیده واگرایی ناشی از خروج از توزیع طول توالی (OOD RoPE issues) به کلی حذف شده است [۵۱].

تحلیل آماری و مدل‌سازی رفتار ریاضی دادگان طولانی (Statistical Modeling & Power Law) کارایی دادگان طولانی از طریق اندازه‌گیری میانگین تجمعی لگاریتم منفی درست‌نمایی (Cumulative Average NLL) به عنوان تابعی از موقعیت توکن (t) روی کدهای بزرگ‌تر از ۱ میلیون توکن ارزیابی شده است:

$$\overline{\mathsf{NLL}}(t) = -\frac{1}{t}\sum_{i=1}^{t}\log P(x_i \mid x_1, x_2, \ldots, x_{i-1}) \tag{۲۱}$$

توزیع داده‌ها و برازش ریاضی نشان می‌دهد که رابطه میان طول دنباله ورودی و افت خطای پیش‌بینی، به دقت از یک رابطه قانون توان (Power Law) تبعیت می‌کند:

$$\overline{\mathsf{NLL}}(t) \approx \alpha \cdot t^{-\beta} + \gamma \tag{۲۲}$$

برازش رگرسیونی این رابطه روی داده‌های کد به ضریب تبیین فوق‌العاده $R^2 = 0.883$ دست یافته است. مقدار $\overline{\mathsf{NLL}}$ از تراز اولیه ≈ 1.2 در فواصل ابتدایی ($t \approx 128$) با روندی پیوسته به مجاورت ≈ 0.4 در فواصل انتهایی ($t \approx 1$M) تقلیل می‌یابد که دال بر توانایی مدل در بهره‌برداری مؤثر از داده‌های پس‌زمینه در سراسر پنجره ۱ میلیون توکنی است.

قالب‌بندی عددی دادگان و پایداری در کوانتایزیشن NVFP4 آموزش تانسورهای داده‌ای بر روی فرمت NVFP4 با بهره‌گیری از میکروبلوک‌های ۱۶ عنصری، فاکتور مقیاس‌بندی E4M3 و مقیاس سراسری FP32 انجام شد [۵۲].

- در مدل Nano با ۳ میلیارد پارامتر فعال (A3B)، شکاف اتلاف داده‌ها بین آموزش NVFP4 و فرمت BF16 به کمتر از **۱٪** تقلیل یافت.
- در مدل‌های بزرگ‌تر با ۸ میلیارد پارامتر فعال (A8B)، شکاف اتلاف به کمتر از **۰.۶٪** محدود گردید که نشان‌دهنده کاهش خطای داده‌ها با افزایش اندازه مدل مطابق با قوانین مقیاس‌پذیری کوانتایزیشن است [۵۳].

ارزیابی‌های آماری مدل‌های پایه بر روی دادگان پیش‌آموزش (Base Model Evaluations) کیفیت پیکره و معماری داده‌های پیش‌آموزش در قالب جداول زیر بررسی شده است. جدول ۱۲ عملکرد داده‌ای معماری LatentMoE را در مقایسه با ساختار استاندارد روی ۱ تریلیون توکن نشان می‌دهد. همچنین جدول ۱۳ نشان‌دهنده ارتقای ۴.۲ درصدی بهره‌وری از دادگان در صورت بهره‌گیری از ماژول پیش‌بینی چندتوکنی (MTP) [۲۷] است. در نهایت، جدول ۱۴ تاب‌آوری استخراج داده در بافت‌های فوق‌طولانی بنچمارک RULER [۳۸] را پس از فاز CPT گزارش می‌کند.

جدول ۱۲: مقایسه دقت بر روی داده‌های ارزیابی پس از آموزش روی ۱ تریلیون توکن بین Standard MoE و LatentMoE

معماری مدل	MMLU-Pro (%)	MMLU (%)	کد (Code) (%)	ریاضی (Math) (%)	درک شهودی (Commonsense) (%)
Standard MoE (8.09B active / 72.6B total)	۳۰.۴۸	۱۰.۷۰	۹۵.۵۱	۳۲.۷۸	۷۳.۸۱
LatentMoE (8.02B active / 72.8B total)	**۸۷.۵۲**	**۱۱.۷۲**	**۱۴.۵۵**	**۱۹.۸۰**	**۱۰.۸۲**

جدول ۱۳: ارزیابی تأثیر نظارت متراکم دادگان با پیش‌بینی چندتوکنی (MTP) بر مدل پایه 8B MoE (آموزش روی ۱ تریلیون توکن)

دسته و بنچمارک ارزیابی	مدل پایه (8B MoE Base)	پایه به همراه MTP (Base + MTP)
دانش عمومی (General Knowledge)		
MMLU (5-shot, acc)	۰۶.۷۰	**۲۶.۷۱**
MMLU-Pro (5-shot, CoT EM)	۰۵.۴۵	**۸۴.۴۷**
کدنویسی (Code)		
MBPP-Sanitized (3-shot)	۵۸.۶۵	**۸۹.۶۶**
درک عمومی (Commonsense Understanding)		
ARC-Challenge (25-shot, acc_norm)	۴۳.۸۶	**۰۵.۸۸**
WinoGrande (0-shot, acc)	۵۹.۷۴	**۴۵.۷۵**
درک مطلب (Reading Comprehension)		
RACE (0-shot, acc)	۰۲.۸۴	**۳۶.۸۵**
ریاضیات (Math)		
GSM8K (8-shot, acc)	۴۹.۸۲	**۴۶.۸۴**

جدول ۱۴: نتایج محک RULER جهت سنجش اکستراپولاسیون بافت پس از فاز CPT در طول توالی 512K

مدل	۱۲۸k	۲۵۶k	۵۱۲k	۱M
(Dense Hybrid) **Nemotron-Nano-12B-v2-Base**	۱۳.۸۵	۸۵.۷۹	۱۲.۷۵	۴۳.۲۳
(MoE Hybrid) **Nemotron-3-Nano-30B-A3B-Base**	۴۸.۷۴	۶۷.۷۱	۰۲.۶۶	**۱۹.۵۴**

حاکمیت داده، شفافیت و عوامل مشارکت‌کننده انویدیا راهبردهای مهندسی داده، پایپ‌لاین‌های آماده‌سازی و بیش از ۱۰ تریلیون توکن از پیکره آموزشی را به شکل عمومی آزاد کرده است. طراحی و فیلترینگ این مجموعه داده عظیم توسط تیم توسعه داده‌های پیش‌آموزش شامل Abhinav Khattar،

Aleksander Ficek، Alisa Liu، Brandon Norick، Eileen Long، Mohammad Shoeybi و بیش از ۵۰ پژوهشگر ارشد دیگر هدایت شده است.

- **Qwen3**

مقدمه و ساختار کلی دادگان پیش‌آموزش فاز پیش‌آموزش (Pre-training) در مدل‌های زبانی خانواده **Qwen3** با هدف جهش در کارایی محاسباتی، تعمیق قابلیت‌های استدلال ریاضی و برنامه‌نویسی و ارتقای شگرف پوشش چندزبانه طراحی شده است. در این نسخه، مدل‌ها بر روی مقیاس بی‌سابقه‌ای بالغ بر **۳۶ تریلیون توکن** (36T tokens) آموزش دیده‌اند که نمایانگر رشد دو برابری نسبت به نسل پیشین (Qwen2.5) است:

$$\rho_{\text{tokens}} = \frac{T_{\text{Qwen۳}}}{T_{\text{Qwen۲.۵}}} = \frac{36 \times 10^{12}}{18 \times 10^{12}} = 2.0 \tag{۲۳}$$

همچنین گستره زبانی دادگان از ۲۹ زبان در نسخه پیشین به **۱۱۹ زبان و گویش** بسط یافته است که حاکی از افزایش بیش از چهار برابری در تنوع زبان‌شناختی است:

$$\rho_{\text{lang}} = \frac{N_{\text{lang}}^{\text{Qwen۳}}}{N_{\text{lang}}^{\text{Qwen۲.۵}}} = \frac{119}{29} \approx 4.103 \tag{۲۴}$$

پیکره پیش‌آموزش به‌گونه‌ای توازن یافته است که از ۶ مدل متراکم (Dense) در مقیاس‌های 0.6B، 1.7B، 4B، 8B، 14B و 32B و دو مدل مبتنی بر ترکیب متخصصان (MoE) شامل Qwen3-30B-A3B (با ۳ میلیارد پارامتر فعال) و مدل پرچم‌دار Qwen3-235B-A22B (دارای ۲۳۵ میلیارد پارامتر کل و ۲۲ میلیارد پارامتر فعال به ازای هر توکن) پشتیبانی نماید. هدف بهینه‌سازی در این مرحله، کمینه‌سازی خطای پیش‌بینی خودبازگشتی توکن بعدی بر روی توالی‌های متنی است:

$$\mathcal{L}_{\text{pretrain}}(\theta) = -\sum_{t=1}^{T} \log P_\theta(x_t \mid x_1, x_2, \ldots, x_{t-1}) \tag{۲۵}$$

گردآوری، استخراج چندوجهی و سنتز هدفمند داده‌ها (Data Sourcing & Synthesis) برای تحقق پیکره ۳۶ تریلیون توکنی با چگالی دانشی بالا، از یک خط لوله سه‌گانه و ترکیبی بهره گرفته شده است:

- **استخراج متن اسنادی بر پایه مدل‌های چندوجهی (Multimodal Document Extraction):** بخش وسیعی از اسناد علمی، دانشگاهی و کتاب‌ها در قالب پرونده‌های PDF متراکم محبوس مانده‌اند. برای بازیابی دقیق و ساختاریافته متون، ابتدا مدل بینایی-زبانی Qwen2.5-VL [۵۴] بازتنظیم شد تا نویسه‌خوانی و درک چیدمان بصری اسناد را با دقت بالا انجام دهد. سپس خروجی حاصل با استفاده از مدل Qwen2.5 [۵۵] بازبینی و پالایش زبانی شد. این فرایند دومرحله‌ای به آزادسازی تریلیون‌ها توکن متنی باکیفیت و با ساختار صحیح انجامید.
- **سنتز داده‌های تخصصی توسط مدل‌های خبره (Domain-Specific Synthetic Data):** به منظور تعمیق توانایی‌های استدلالی، تریلیون‌ها توکن سنتزی در قالب کتاب‌های درسی، جفت‌های پرسش‌وپاسخ استدلالی، دستورالعمل‌ها و قطعه‌کدهای برنامه‌نویسی به مجموعه افزوده شد. در این مسیر، داده‌های ریاضی به طور ویژه توسط مدل تخصصی Qwen2.5-Math [۵۶] و کدهای چندزبانه به همراه تحلیل الگوریتمی توسط مدل Qwen2.5-Coder [۵۷] سنتز گردیدند.
- **توسعه چندزبانه متوازن (Multilingual Expansion):** پوشش زبانی از ۲۹ به ۱۱۹ زبان گسترش یافت؛ داده‌های زبان‌های دارای منابع متوسط و کم با راهبردهای ترجمه معکوس و اعتبارسنجی کیفی تقویت شدند تا نمایش چندزبانه در لایه‌های بازنمایی مدل متوازن گردد.

حاشیه‌نویسی چندزبانه و بهینه‌سازی توزیع داده در سطح نمونه (Instance-Level Data Mixture) مدیریت و مهندسی این حجم از داده نیازمند گزینش کیفی فراتر از فیلترینگ‌های سنتی است:

- **سامانه حاشیه‌نویسی چندبعدی:** یک سیستم خودکار برچسب‌گذاری بر روی بیش از **۳۰ تریلیون توکن** اعمال شد. این سامانه داده‌ها را بر اساس چهار بعد کلیدی ارزیابی و برچسب‌گذاری می‌کند: ارزش آموزشی (Educational Value)، شاخه علمی (Field)، دامنه تخصصی (Domain) و انطباق با شاخص‌های ایمنی (Safety).
- **بهینه‌سازی توزیع در سطح نمونه (Instance-Level Optimization):** برخلاف رویکردهای پیشین نظیر DoREMI [۵۸]، DoGE [۵۹] و RegMix [۶۰] که وزن‌دهی به داده‌ها را در سطح منبع داده یا حوزه کلی بهینه‌سازی می‌کردند، در **Qwen3** ترکیب داده‌ها بر پایه برچسب‌های ریزدانه در سطح تک‌تک نمونه‌ها تنظیم شد. این فرایند با اجرای مطالعات ابلیشن گسترده بر روی مدل‌های پروکسی کوچک (Small Proxy Models) انجام گرفت تا بردار ضرایب ترکیب بهینه w^* حاصل شود:

$$w^* = \arg\min_{w} \mathbb{E}_{x\sim\mathcal{D}_{\text{eval}}} \left[\mathcal{L}_{\text{proxy}}(x;\theta^*(w))\right] \tag{۲۶}$$

برنامه درسی سه‌مرحله‌ای پیش‌آموزش (Three-Stage Pre-training Curriculum) فرآیند پیش‌آموزش در قالب سه فاز زمان‌بندی‌شده و ساختاریافته به اجرا درآمد:

- **فاز ۱: مرحله عمومی (General Stage - S1):** آموزش بر روی بیش از ۳۰ تریلیون توکن با طول زمینه $L = 4{,}096$ توکن اجرا شد. این مرحله پایه دانش زبانی، استنتاج عمومی جهان و تعمیم‌پذیری متقابل در ۱۱۹ زبان و گویش را شکل داد.
- **فاز ۲: مرحله استدلال (Reasoning Stage - S2):** مدل‌ها با حدود ۵ تریلیون توکن گلچین‌شده و بسیار غنی با محوریت علوم مهندسی، فناوری، ریاضیات (STEM)، مسائل الگوریتمی و داده‌های سنتزی با طول توالی $4{,}096$ آموزش دیدند. در این فاز، نرخ اضمحلال یادگیری (Learning Rate Decay) تسریع گردید تا مدل در فضاهای استدلالی همگرا شود.
- **فاز ۳: مرحله زمینه طولانی (Long Context Stage):** مدل‌ها با صدها میلیارد توکن بلند تا سقف طول توالی $L = 32{,}768$ توکن پیش‌آموزش دیدند. ساختار طول اسناد در این فاز شامل **۷۵٪ داده‌ها با طول بین ۱۶،۳۸۴ تا ۳۲،۷۶۸ توکن** و **۲۵٪ داده‌ها با طول بین ۴،۰۹۶ تا ۱۶،۳۸۴ توکن** بوده است. در این فاز بسامد پایه فشرده‌سازی موقعیتی دورانی (RoPE) [۳۶] با تکنیک ABF [۶۱] از ۱۰،۰۰۰ به ۱،۰۰۰،۰۰۰ افزایش یافت و با سازوکارهای YaRN [۱۷] و توجه چندتکه‌ای دوگانه (DCA) [۶۲] ادغام شد تا پنجره بافت حین استنتاج تا سقف ۱۲۸،۰۰۰ توکن (128K) با حفظ پایایی بسط یابد.

توکنایزیشن و بازنمایی زیرکلمات (Tokenization) در مدل‌های **Qwen3** از توکنایزر اختصاصی مبتنی بر کدگذاری جفت‌بایت در سطح بایت (BBPE) [۶۳، ۶۴، ۶۵، ۶۶] با اندازه واژگان $151{,}669$ استفاده شده است. این اندازه واژگان، ضریب فشرده‌سازی توکن به کلمه را در متون غیرانگلیسی به خصوص خطوط خاورمیانه و آسیایی بهینه ساخته و سربار محاسباتی طول توالی را تعدیل می‌کند.

ارزیابی‌های آماری و مقایسه‌ای مدل‌های پایه (Base Model Evaluations) اثربخشی داده‌های پیش‌آموزش و راهبردهای پالایش آن، بر روی ۱۵ بنچمارک استاندارد در ۴ حوزه وظایف عمومی (MMLU [۳۰]، MMLU-Redux [۶۷]، MMLU-Pro [۶۸]، BBH [۶۹]، SuperGPQA [۷۰])، ریاضیات و علوم پایه (GPQA [۷۱]، GSM8K [۷۲]، MATH [۲۹])، کدنویسی (EvalPlus [۷۳]، MultiPL-E [۷۴]، MBPP [۷۵]، CRUX-O [۷۶]) و چندزبانه (MGSM [۷۷]، MMMLU [۷۸]، INCLUDE [۷۹]) مورد ارزیابی دقیق قرار گرفت.

جداول ۱۵ تا ۲۰ عملکرد کمّی مدل‌های پایه را در برابر رقبای هم‌رده و پرچم‌دار نشان می‌دهند. مهم‌ترین یافته‌های آماری به شرح زیر است:

- **برتری مدل پرچم‌دار Qwen3-235B-A22B-Base:** این مدل با دارا بودن تنها ۲۲ میلیارد پارامتر فعال در هر توکن (تقریباً 60% پارامترهای فعال و 35% پارامترهای کل DeepSeek-V3 [۸۰])، در ۱۴ آزمون از ۱۵ آزمون ارزیابی‌شده رتبه اول را کسب کرده است؛ از جمله در آزمون استدلال پیشرفته MATH با ثبت امتیاز 71.84 در مقایسه با 62.62 مدل DeepSeek-V3، رشد نسبی $+14.72\%$ را محقق ساخته است.
- **انتقال تراز کارایی در مدل‌های متراکم (Dense Models):** مدل Qwen3-32B-Base با کمتر از نیمی از پارامترهای مدل پرچم‌دار پیشین Qwen2.5-72B-Base، در ۱۰ بنچمارک از ۱۵ بنچمارک بر آن غلبه کرده و در کدنویسی (EvalPlus با امتیاز 72.05 در برابر 65.93) و آزمون MMLU-Pro (امتیاز 65.54 در برابر 58.07) برتری قاطعی به ثبت رسانده است. همچنین مدل‌های سبک‌تر Qwen3-8B، 4B و 1.7B توانسته‌اند مدل‌های مقیاس‌های ۲ تا ۳ برابری نسل قبل خود (Qwen2.5-14B، 7B و 3B) را در اکثر حوزه‌ها پشت سر بگذارند.
- **بهره‌وری فوق‌العاده مدل‌های MoE کوچک:** مدل Qwen3-30B-A3B-Base با تنها ۳ میلیارد پارامتر فعال، در تمامی شاخص‌ها مدل متراکم Qwen2.5-14B-Base (با ۱۴ میلیارد پارامتر فعال) را مغلوب کرده و به کارایی نزدیک به مدل متراکم ۳۲ میلیارد پارامتری با بیش از ۱۰ برابر هزینه محاسباتی کمتر دست یافته است.

جدول ۱۵: مقایسه عملکرد مدل پایه پرچم‌دار Qwen3-235B-A22B-Base با مدل‌های قدرتمند متن‌باز

شاخص / بنچمارک	Qwen2.5-72B-Base	Qwen2.5-Plus-Base	Llama-4-Maverick-Base	DeepSeek-V3-Base	Qwen3-235B-A22B-Base
معماری	Dense	MoE	MoE	MoE	MoE
تعداد کل پارامترها	۷۲B	۲۷۱B	۴۰۲B	۶۷۱B	۲۳۵B
پارامترهای فعال	۷۲B	۳۷B	۱۷B	۳۷B	۲۲B
وظایف عمومی (General Tasks)					
MMLU	۰۶.۸۶	۰۲.۸۵	۱۶.۸۵	۱۹.۸۷	**۸۱.۸۷**
MMLU-Redux	۹۱.۸۳	۶۹.۸۲	۰۵.۸۴	۱۴.۸۶	**۴۰.۸۷**
MMLU-Pro	۰۷.۵۸	۵۲.۶۳	۹۱.۶۳	۸۴.۵۹	**۱۸.۶۸**
SuperGPQA	۲۰.۳۶	۱۸.۳۷	۸۵.۴۰	۵۳.۴۱	**۰۶.۴۴**
BBH	۳۰.۸۶	۶۰.۸۵	۶۲.۸۳	۲۲.۸۶	**۸۷.۸۸**
ریاضیات و علوم مهندسی (Math & STEM)					
GPQA	۸۸.۴۵	۹۲.۴۱	۹۴.۴۳	۹۲.۴۱	**۴۷.۴۷**
GSM8K	۵۰.۹۱	۸۹.۹۱	۷۲.۸۷	۵۷.۸۷	**۳۹.۹۴**
MATH	۱۲.۶۲	۷۸.۶۲	۳۲.۶۳	۶۲.۶۲	**۸۴.۷۱**
کدنویسی (Coding)					
EvalPlus	۹۳.۶۵	۴۳.۶۱	۳۸.۶۸	۷۵.۶۳	**۶۰.۷۷**
MultiPL-E	۷۰.۵۸	۱۶.۶۲	۲۸.۵۷	۲۶.۶۲	**۹۴.۶۵**
MBPP	۰۰.۷۶	۶۰.۷۴	۴۰.۷۵	۲۰.۷۴	**۴۰.۸۱**
CRUX-O	۲۰.۶۶	۵۰.۶۸	۰۰.۷۷	۶۰.۷۶	**۰۰.۷۹**
چندزبانه (Multilingual)					
MGSM	۴۰.۸۲	۲۱.۸۲	۶۹.۷۹	۶۸.۸۲	**۵۳.۸۳**
MMMLU	۴۰.۸۴	۴۹.۸۳	۰۹.۸۳	۸۸.۸۵	**۷۰.۸۶**
INCLUDE	۰۵.۶۹	۹۷.۶۶	۴۷.۷۳	**۱۷.۷۵**	۴۶.۷۳

- **OLMo 3**

مقدمه و فلسفه بنیادین دادگان پیش‌آموزش (Pre-Training Data Philosophy) در چارچوب چرخه کامل و شفاف مدل‌های زبانی سری **OLMo 3** (Model Flow)، فاز پیش‌آموزش با هدف برپایی زیربنایی مستحکم جهت شکوفایی قابلیت‌های استدلال عمیق گام‌به‌گام (Thinking) و تعامل با ابزارها (Tool Use) در مراحل پس‌آموزش طراحی شده است [۸۱]. به این منظور، پیکره عظیم **Dolma 3** در قالب یک استخر داده جامع به وسعت **۳۱.۹ تریلیون توکن** (شامل ۹۷.۹ میلیارد سند) گردآوری شده و ترکیب بهینه‌سازی‌شده نهایی آن موسوم به **Dolma 3 Mix** با حجم **۹۳.۵ تریلیون توکن** (شامل ۸۷.۳ میلیارد سند) مبنای پیش‌آموزش مدل‌های پایه در دو مقیاس ۷ و ۳۲ میلیارد پارامتر قرار گرفته است.

با توجه به این‌که بیش از ۹۰ درصد کل توان محاسباتی چرخه حیات مدل در فاز پیش‌آموزش مصرف می‌شود، استراتژی تدوین این پیکره بر دو اصل راهبردی استوار است:

جدول ۱۶: مقایسه عملکرد مدل پایه متراکم Qwen3-32B-Base با مدل‌های هم‌رده

شاخص / بنچمارک	Qwen2.5-32B-Base	Qwen2.5-72B-Base	Gemma-3-27B-Base	Llama-4-Scout-Base	Qwen3-32B-Base
معماری	Dense	Dense	Dense	MoE	Dense
تعداد کل پارامترها	۳۲B	۷۲B	۲۷B	۱۰۹B	۳۲B
پارامترهای فعال	۳۲B	۷۲B	۲۷B	۱۷B	۳۲B
وظایف عمومی (General Tasks)					
MMLU	۳۲.۸۳	**۰۶.۸۶**	۶۹.۷۸	۲۷.۷۸	۶۱.۸۳
MMLU-Redux	۹۷.۸۱	**۹۱.۸۳**	۵۳.۷۶	۰۹.۷۱	۴۱.۸۳
MMLU-Pro	۱۰.۵۵	۰۷.۵۸	۸۸.۵۲	۱۳.۵۶	**۵۴.۶۵**
SuperGPQA	۵۵.۳۳	۲۰.۳۶	۸۷.۲۹	۵۱.۲۶	**۷۸.۳۹**
BBH	۴۸.۸۴	۳۰.۸۶	۹۵.۷۹	۴۰.۸۲	**۳۸.۸۷**
ریاضیات و علوم مهندسی (Math & STEM)					
GPQA	۹۷.۴۷	۸۸.۴۵	۲۶.۲۶	۴۰.۴۰	**۴۹.۴۹**
GSM8K	۸۷.۹۲	۵۰.۹۱	۲۰.۸۱	۳۷.۸۵	**۴۰.۹۳**
MATH	۷۰.۵۷	**۱۲.۶۲**	۷۸.۵۱	۶۶.۵۱	۶۲.۶۱
کدنویسی (Coding)					
EvalPlus	۲۵.۶۶	۹۳.۶۵	۷۸.۵۵	۹۰.۵۹	**۰۵.۷۲**
MultiPL-E	۳۰.۵۸	۷۰.۵۸	۰۳.۴۵	۳۸.۴۷	**۰۶.۶۷**
MBPP	۶۰.۷۳	۰۰.۷۶	۴۰.۶۸	۶۰.۶۸	**۲۰.۷۸**
CRUX-O	۸۰.۶۷	۲۰.۶۶	۰۰.۶۰	۹۰.۶۱	**۵۰.۷۲**
چندزبانه (Multilingual)					
MGSM	۱۲.۷۸	۴۰.۸۲	۷۴.۷۳	۹۳.۷۹	**۰۶.۸۳**
MMMLU	۴۰.۸۲	**۴۰.۸۴**	۶۲.۷۷	۸۳.۷۴	۸۳.۸۳
INCLUDE	۳۵.۶۴	**۰۵.۶۹**	۹۴.۶۸	۰۹.۶۸	۸۷.۶۷

جدول ۱۷: مقایسه عملکرد مدل‌های Qwen3-14B-Base و Qwen3-30B-A3B-Base با مدل‌های هم‌رده

شاخص / بنچمارک	Gemma-3-12B	Qwen2.5-14B	Qwen2.5-32B	Qwen2.5-Turbo	Qwen3-14B	Qwen3-30B-A3B
معماری	Dense	Dense	Dense	MoE	Dense	MoE
تعداد کل پارامترها	۱۲B	۱۴B	۳۲B	۴۲B	۱۴B	۳۰B
پارامترهای فعال	۱۲B	۱۴B	۳۲B	۶B	۱۴B	۳B
وظایف عمومی (General Tasks)						
MMLU	۸۷.۷۳	۶۶.۷۹	**۳۲.۸۳**	۵۰.۷۹	۰۵.۸۱	۳۸.۸۱
MMLU-Redux	۷۰.۷۰	۶۴.۷۶	**۹۷.۸۱**	۱۱.۷۷	۸۸.۷۹	۱۷.۸۱
MMLU-Pro	۹۱.۴۴	۱۶.۵۱	۱۰.۵۵	۶۰.۵۵	۰۳.۶۱	**۴۹.۶۱**
SuperGPQA	۶۱.۲۴	۶۸.۳۰	۵۵.۳۳	۱۹.۳۱	۲۷.۳۴	**۷۲.۳۵**
BBH	۲۸.۷۴	۱۸.۷۸	**۴۸.۸۴**	۱۰.۷۶	۰۷.۸۱	۵۴.۸۱
ریاضیات و علوم مهندسی (Math & STEM)						
GPQA	۳۱.۳۱	۸۳.۳۲	**۹۷.۴۷**	۴۱.۴۱	۹۰.۳۹	۹۴.۴۳
GSM8K	۰۱.۷۸	۲۲.۹۰	**۸۷.۹۲**	۳۲.۸۸	۴۹.۹۲	۸۱.۹۱
MATH	۴۳.۴۴	۶۴.۵۵	۷۰.۵۷	۶۰.۵۵	**۰۲.۶۲**	۰۴.۵۹
کدنویسی (Coding)						
EvalPlus	۶۵.۵۲	۷۰.۶۰	۲۵.۶۶	۲۳.۶۱	**۲۳.۷۲**	۴۵.۷۱
MultiPL-E	۰۳.۴۳	۷۹.۵۴	۳۰.۵۸	۲۴.۵۳	۶۹.۶۱	**۵۳.۶۶**
MBPP	۶۰.۶۰	۰۰.۶۹	۶۰.۷۳	۶۰.۶۷	۴۰.۷۳	**۴۰.۷۴**
CRUX-O	۰۰.۵۲	۱۰.۶۱	۸۰.۶۷	۲۰.۶۰	**۶۰.۶۸**	۲۰.۶۷
چندزبانه (Multilingual)						
MGSM	۳۵.۶۴	۶۸.۷۴	۱۲.۷۸	۴۵.۷۰	**۲۰.۷۹**	۱۱.۷۹
MMMLU	۵۰.۷۲	۳۴.۷۸	**۴۰.۸۲**	۷۶.۷۹	۶۹.۷۹	۴۶.۸۱
INCLUDE	۳۴.۶۳	۲۶.۶۰	۳۵.۶۴	۲۵.۵۹	۵۵.۶۴	**۰۰.۶۷**

جدول ۱۸: مقایسه عملکرد مدل پایه Qwen3-8B-Base با مدل‌های هم‌رده

شاخص / بنچمارک	Llama-3-8B-Base	Qwen2.5-7B-Base	Qwen2.5-14B-Base	Qwen3-8B-Base
معماری	Dense	Dense	Dense	Dense
تعداد کل پارامترها	۸B	۷B	۱۴B	۸B
پارامترهای فعال	۸B	۷B	۱۴B	۸B
وظایف عمومی (General Tasks)				
MMLU	۶۰.۶۶	۱۶.۷۴	**۶۶.۷۹**	۸۹.۷۶
MMLU-Redux	۵۹.۶۱	۰۶.۷۱	**۶۴.۷۶**	۱۷.۷۶
MMLU-Pro	۳۶.۳۵	۰۰.۴۵	۱۶.۵۱	**۷۳.۵۶**
SuperGPQA	۵۴.۲۰	۳۴.۲۶	۶۸.۳۰	**۶۴.۳۱**
BBH	۷۰.۵۷	۴۰.۷۰	۱۸.۷۸	**۴۰.۷۸**
ریاضیات و علوم مهندسی (Math & STEM)				
GPQA	۸۰.۲۵	۳۶.۳۶	۸۳.۳۲	**۴۴.۴۴**
GSM8K	۳۰.۵۵	۳۶.۸۵	**۲۲.۹۰**	۸۴.۸۹
MATH	۵۰.۲۰	۸۰.۴۹	۶۴.۵۵	**۸۰.۶۰**
کدنویسی (Coding)				
EvalPlus	۱۳.۴۴	۱۸.۶۲	۷۰.۶۰	**۶۵.۶۷**
MultiPL-E	۴۵.۳۱	۷۳.۵۰	۷۹.۵۴	**۷۵.۵۸**
MBPP	۴۰.۴۸	۴۰.۶۳	۰۰.۶۹	**۸۰.۶۹**
CRUX-O	۸۰.۳۶	۵۰.۴۸	۱۰.۶۱	**۰۰.۶۲**
چندزبانه (Multilingual)				
MGSM	۹۲.۳۸	۶۰.۶۳	۶۸.۷۴	**۰۲.۷۶**
MMMLU	۶۵.۵۹	۳۴.۷۱	**۳۴.۷۸**	۷۲.۷۵
INCLUDE	۹۴.۴۴	۹۸.۵۳	**۲۶.۶۰**	۴۰.۵۹

جدول ۱۹: مقایسه عملکرد مدل پایه سبک Qwen3-4B-Base با مدل‌های هم‌رده

شاخص / بنچمارک	Gemma-3-4B-Base	Qwen2.5-3B-Base	Qwen2.5-7B-Base	Qwen3-4B-Base
معماری	Dense	Dense	Dense	Dense
تعداد کل پارامترها	۴B	۳B	۷B	۴B
پارامترهای فعال	۴B	۳B	۷B	۴B
وظایف عمومی (General Tasks)				
MMLU	۵۱.۵۹	۶۲.۶۵	**۱۶.۷۴**	۹۹.۷۲
MMLU-Redux	۹۱.۵۶	۶۸.۶۳	۰۶.۷۱	**۷۹.۷۲**
MMLU-Pro	۲۳.۲۹	۶۱.۳۴	۰۰.۴۵	**۵۸.۵۰**
SuperGPQA	۶۸.۱۷	۳۱.۲۰	۳۴.۲۶	**۴۳.۲۸**
BBH	۷۰.۵۱	۳۰.۵۶	۴۰.۷۰	**۵۹.۷۲**
ریاضیات و علوم مهندسی (Math & STEM)				
GPQA	۲۴.۲۴	۲۶.۲۶	۳۶.۳۶	**۸۷.۳۶**
GSM8K	۹۷.۴۳	۰۸.۷۹	۳۶.۸۵	**۷۹.۸۷**
MATH	۱۰.۲۶	۶۴.۴۲	۸۰.۴۹	**۱۰.۵۴**
کدنویسی (Coding)				
EvalPlus	۲۳.۴۳	۲۸.۴۶	۱۸.۶۲	**۵۳.۶۳**
MultiPL-E	۰۶.۲۸	۶۵.۳۹	۷۳.۵۰	**۱۳.۵۳**
MBPP	۴۰.۴۶	۶۰.۵۴	۴۰.۶۳	**۰۰.۶۷**
CRUX-O	۰۰.۳۴	۵۰.۳۶	۵۰.۴۸	**۰۰.۵۵**
چندزبانه (Multilingual)				
MGSM	۱۱.۳۳	۵۳.۴۷	۶۰.۶۳	**۷۴.۶۷**
MMMLU	۶۲.۵۹	۵۵.۶۵	۳۴.۷۱	**۴۲.۷۱**
INCLUDE	۰۶.۴۹	۹۰.۴۵	۹۸.۵۳	**۲۹.۵۶**

جدول ۲۰: مقایسه عملکرد مدل‌های لبه‌ای Qwen3-1.7B-Base و Qwen3-0.6B-Base با مدل‌های هم‌رده

شاخص / بنچمارک	Qwen2.5-0.5B	Qwen3-0.6B	Gemma-3-1B	Qwen2.5-1.5B	Qwen3-1.7B
معماری	Dense	Dense	Dense	Dense	Dense
تعداد کل پارامترها	۵B.۰	۶B.۰	۱B	۵B.۱	۷B.۱
پارامترهای فعال	۵B.۰	۶B.۰	۱B	۵B.۱	۷B.۱
وظایف عمومی (General Tasks)					
MMLU	۵۰.۴۷	۸۱.۵۲	۲۶.۲۶	۹۰.۶۰	**۶۳.۶۲**
MMLU-Redux	۱۰.۴۵	۲۶.۵۱	۹۹.۲۵	۴۶.۵۸	**۶۶.۶۱**
MMLU-Pro	۶۹.۱۵	۷۴.۲۴	۷۲.۹	۵۳.۲۸	**۷۶.۳۶**
SuperGPQA	۳۰.۱۱	۰۳.۱۵	۱۹.۷	۶۴.۱۷	**۹۲.۲۰**
BBH	۳۰.۲۰	۴۷.۴۱	۱۳.۲۸	۱۰.۴۵	**۴۷.۵۴**
ریاضیات و علوم مهندسی (Math & STEM)					
GPQA	۷۵.۲۴	۷۷.۲۶	۷۵.۲۴	۲۴.۲۴	**۲۸.۲۸**
GSM8K	۶۲.۴۱	۵۹.۵۹	۲۰.۲	۵۴.۶۸	**۴۴.۷۵**
MATH	۴۸.۱۹	۴۴.۳۲	۶۶.۳	۰۰.۳۵	**۵۰.۴۳**
کدنویسی (Coding)					
EvalPlus	۸۵.۳۱	۲۳.۳۶	۹۸.۸	۸۰.۴۴	**۷۰.۵۲**
MultiPL-E	۷۰.۱۸	۵۸.۲۴	۱۵.۵	۱۰.۳۳	**۷۱.۴۲**
MBPP	۸۰.۲۹	۶۰.۳۶	۲۰.۹	۶۰.۴۳	**۴۰.۵۵**
CRUX-O	۱۰.۱۲	۰۰.۲۷	۸۰.۳	۶۰.۲۹	**۴۰.۳۶**
چندزبانه (Multilingual)					
MGSM	۰۷.۱۲	۹۹.۳۰	۷۴.۱	۸۲.۳۲	**۷۱.۵۰**
MMMLU	۵۳.۳۱	۱۶.۵۰	۵۷.۲۶	۲۷.۶۰	**۲۷.۶۳**
INCLUDE	۷۴.۲۴	۲۶.۳۴	۶۲.۲۵	۵۵.۳۹	**۵۷.۴۵**

- **کفایت مقیاس داده‌ها (Scale Sufficiency):** یک منبع داده تنها در صورتی شایسته ورود به فاز پیش‌آموزش تلقی می‌شود که پتانسیل تولید حجم کافی از توکن‌ها را برای اثرگذاری در مقیاس تریلیونی دارا باشد؛ داده‌های کیفی ولی کم‌حجم به‌طور کامل برای مرحله میان‌آموزش (Mid-training) ذخیره می‌شوند.
- **پرهیز از داده‌های تکلیفی ساختاریافته (Exclusion of Task-Structured Data):** برای جلوگیری از مخدوش شدن نتایج ابلیشن و سوگیری‌های زودرس ارزیابی، داده‌های جفت پرسش‌وپاسخ (QA) و مکالمات دستوری از مرحله پیش‌آموزش حذف و به مراحل بعدی موکول شده‌اند.

ترکیب‌بندی و مشخصات آماری پیکره پیش‌آموزش (Dolma 3 Mix Composition) پیکره متنی پیش‌آموزش Dolma 3 Mix از شش منبع کلیدی به شرح جدول ۲۱ استخراج، پالایش و متوازن شده است. در این فرآیند، استخر اولیه (9T Pool) با اعمال راهبردهای نمونه‌برداری به ترکیب نهایی (6T Mix) و همچنین یک نسخه سبک جهت پژوهش‌های مقیاس‌کوچک (150B Mix) تبدیل شده است.

جدول ۲۱: ترکیب‌بندی و مشخصات آماری منابع مجموعه‌داده Dolma 3 در استخر داده و ترکیب‌های آموزشی

منبع داده	نوع محتوا	استخر داده (9T Pool)		ترکیب نهایی (6T Mix)		ترکیب آزمایشی (150B Mix)	
		توکن	اسناد	توکن (درصد)	اسناد	توکن (درصد)	اسناد
Common Crawl	صفحات وب باز	۱۴T.۸	۶۷B.۹	۵۱T.۴ (۱%.۷۶)	۱۵B.۳	۱۲۱B (۹%.۷۶)	۵M.۸۴
olmOCR Science PDFs	اسناد علمی و مقالات	۹۷۲B	۱۰۱M	۸۰۵B (۶%.۱۳)	۸M.۸۳	۹B.۱۹ (۶%.۱۲)	۲۵M.۲
StackEdu (Rebalanced)	کدهای برنامه‌نویسی گیت‌هاب	۱۳۷B	۱۶۷M	۴۰۹B (۸۹%.۶)	۵۲۶M	۱B.۱۱ (۰۶%.۷)	۳M.۱۴
arXiv	مقالات علمی ساختاریافته با لاتک	۴B.۲۱	۹۵M.۳	۸B.۵۰ (۸۶%.۰)	۱۰M.۹	۲۹B.۱ (۸۲%.۰)	۲۴۷K
FineMath 3+	متون آموزشی ریاضیات در وب	۱B.۳۴	۴M.۲۱	۱۵۲B (۵۶%.۲)	۵M.۹۵	۱۰B.۴ (۶۰%.۲)	۵۷M.۲
Wikipedia & Wikibooks	متون دانشنامه‌ای مرجع	۶۹B.۳	۶۷M.۶	۵۱B.۲ (۰۴%.۰)	۲۴M.۴	۶M.۶۴ (۰۴%.۰)	۱۱۹K
مجموع کل (Total)	–	**۳۱T.۹**	**۹۷B.۹**	**۹۳T.۵ (۱۰۰%)**	**۸۷B.۳**	**۱۵۷B (۱۰۰%)**	**۱۰۴M**

خط لوله مهندسی و پالایش دادگان وب (Web Data Pipeline) بخش اصلی استخر وب از ۱۰۴ دامپ (Dump) پایگاه Common Crawl (از سال ۲۰۱۳ تا پایان دسامبر ۲۰۲۴) گردآوری شده است. استخراج ساختار متنی و پیاده‌سازی متون عاری از برچسب‌های HTML با استفاده از کتابخانه Resiliparse [۸۲] بر روی فایل‌های WARC انجام گرفت که به حجم خام اولیه **۶.۲۵۲ میلیارد سند** انجامید.

فرایند فیلترینگ اکتشافی (Heuristic Filtering) با اقتباس و بهینه‌سازی خط لوله DCLM [۸۳] در چند لایه پیاده‌سازی شد:

- **فیلتر نشانی‌های وب (URL Filtering):** مسدودسازی دامنه‌ها و لینک‌های غیراخلاقی و بدافزار بر مبنای لیست سیاه مجموعه‌های FineWeb [۸۴] و RefinedWeb [۸۵] (حذف حدود ۱٪ اسناد).
- **فیلترهای اکتشافی متن:** حذف اسناد نامتعارف از حیث طول (بسیار کوتاه یا بسیار بلند)، متون با نسبت پایین کاراکترهای الفبایی-عددی و محتواهای دچار تکرار مفرط درونی.
- **پالایش خطوط و عبارات کلیشه‌ای (Line Modifiers):** حذف سطرها و بندهای آلوده به هرزنامه‌های پرتکرار نظیر «items in cart» یا «read more...» و حذف اسنادی که بخش عمده ساختار خود را در این اصلاح از دست دادند.
- **جداسازی زبانی (FastText Language Filter):** به‌کارگیری مدل طبقه‌بند زبانی با آستانه اطمینان 0.65 برای گزینش زبان انگلیسی.
- **قواعد کیفی MADLAD-400:** اعمال قواعد گزینش ۲ و ۵ برگرفته از پژوهش کودوگونتا و همکاران [۸۶] جهت پالایش جملاتی با تعداد کلمات تمام‌بزرگ یا الگوهای منظم غیرعادی (Cursed regex).

این مراحل در مجموع با صرف ۱۰۳۰ ساعت محاسباتی روی ماشین‌های i4i.32xlarge ابری (معادل ۱۱،۳۰۰ دلار هزینه) منجر به نرخ حذف کلی ۸۴.٪۹ و بقای **۷.۳۸ میلیارد سند** گردید (جدول ۲۲).

جدول ۲۲: ریز تفکیک مراحل پالایش اکتشافی پیکره وب Common Crawl

مرحله پالایش (Pipeline Step)	اسناد حذف‌شده (میلیارد)	درصد حذف از کل استخر	درصد زمان پردازش
فیلترهای URL	۳.۲	۹۰%.۰	۶۸%.۱
فیلترهای طول سند	۴.۱۰۳	۴۲%.۴۰	۰۳%.۸
فیلترهای نماد و تراکم کاراکتر	۵.۵۶	۱۰%.۲۲	۱۳%.۴
تکرار درونی (Internal Repetition)	۱.۳۲	۵۳%.۱۲	۴۱%.۳۱
کننده‌های خطوط (Line Modifiers)	۱.۷	۷۹%.۲	۰۰%.۱۰
فیلتر زبان انگلیسی (FastText)	۲.۶	۴۴%.۲	۳۰%.۱۴
فیلترهای MADLAD-400	۳.۹	۶۵%.۳	۸۷%.۵
طبقه‌بندهای کیفیت	۰.۰	۰۰%.۰	۵۸%.۲۴
مجموع کل (Total)	**۹.۲۱۳**	**۹%.۸۴ حذف (بقای ۱%.۱۵)**	**۱۰۰% (۱،۰۳۰ ساعت)**

معماری سه‌سطحی حذف تکرار مقیاس‌بالا (Three-Stage Deduplication) به منظور افزایش بهره‌وری مصرف توکن [۸۷]، حذف هم‌پوشانی‌های ضعیف به عنوان شاخص کیفیت نادقیق [۸۸]، و پیشگیری از افت کارایی ناشی از تکرار بیش از حد متون [۸۹]، یک راهبرد سه‌سطحی با ابزار اختصاصی **Duplodocus** (توسعه‌یافته به زبان Rust) و **bsade** به کار رفت:

- **حذف تکرار دقیق سراسری (Exact Deduplication):** در دو گام پی‌درپی با بهره‌گیری از هَش متنی ۱۲۸ بیتی؛ ابتدا درون هر یک از ۱۰۴ دامپ (حذف ٪۲۴ اسناد) و سپس به‌طور سراسری در میان کل دامپ‌ها (حذف ٪۴۳ اسناد باقی‌مانده). در نتیجه، ٪۶۶ از اسناد اولیه حذف و استخر به **۷.۱۲ میلیارد سند** کاهش یافت.
- **حذف تکرار فازی در سطح سند (MinHash Fuzzy Deduplication):** افراز داده‌ها به ۳۲ قطعه (Shard)، توکنایزیشن با مدل p50k و استخراج ۵-گِرم‌ها. الگوریتم MinHash LSH با ۲۶ باند به اندازه ۱۱ برای هدف‌گذاری تشابه جاکارد $J \geq 0.80$ پیاده شد. در مرحله بعد، اعتبارسنجی دوبه‌دو برای کنترل مثبت کاذب بر پایه ۳-گِرم‌ها صورت پذیرفت؛ برای خوشه‌های ۵۰۰ سند یا بیشتر، ۲۰۰ باند به اندازه ۳۱ اعمال شد و برای خوشه‌های کوچک‌تر، فاصله جاکارد به صورت دقیق محاسبه گردید. با نگه‌داشتن جدیدترین نسخه سند بر مبنای تاریخ خزش، ٪۲۴ اسناد حذف و استخر به **۸.۹ میلیارد سند** رسید.
- **حذف زیررشته‌های تکراری با آرایه پسوندی (Fuzzy Suffix Array Deduplication):** با استفاده از bsade در ۵۶ بخش، تمام زیررشته‌های دارای طول ۵۰۰ بایت یا بیشتر که حداقل دو بار تکرار شده بودند نشانه‌گذاری شدند. طبق یک قاعده نوآورانه فازی، هرگاه یک بخش متنی از دو سو با توالی‌های تکراری ۵۰۰ بایتی محصور شده بود و بیش از ٪۸۰ آن بخش هم‌پوشانی تکراری داشت، کل توالی حذف شد تا قطعات ریز و ناقص از بین بروند. این گام موجب حذف ٪۱۴ از بایت‌های متنی شده و پیکره نهایی وب را به **۷.۹ میلیارد سند (۵.۳۶ ترابایت متن، معادل ۸ تریلیون توکن)** رساند.

دسته‌بندی موضوعی، کیفی و ساختار ۴۸۰گانه سطل‌ها پس از حذف تکرار، استخر داده‌ها در یک ماتریس دوبعدی از موضوع و کیفیت افراز شد:

- **طبقه‌بندی موضوعی (Topic Classification):** نگاشت متون به ۲۴ رده ساختاریافته برگرفته از WebOrganizer [۹۰]. جهت مقرون‌به‌صرفه بودن پردازش تریلیونی، یک طبقه‌بند FastText بر روی نمونه‌های نشاندار با GPT-4.1 و o4-mini آموزش داده شد که به دقت و فراخوانی متوازن 0.762 در جامعه آماری آزمون دست یافت.
- **ارزیابی کیفی (Quality Tiering):** با استفاده از مدل FastText آموزش‌دیده بر نمونه‌های مثبت شامل OpenHermes-2.5 [۹۱]، ELI5 [۹۲]، UltraChat [۹۳] و WildChat [۹۴]، اسناد به ۲۰ بازه صدکی کیفی ۵ درصدی (Vigintile) تقسیم شدند.

تلاقی ۲۴ رده موضوعی و ۲۰ طبقه کیفی، **۴۸۰ سطل مجزا و متعامد** را شکل داد که دستمایه بهینه‌سازی توزیع در مرحله ترکیب گردید.

پیکره مقالات و اسناد آکادمیک (olmOCR Science PDFs) در یک پیشرفت شاخص نسبت به دادگان peS2o [۹۵]، مدل **OLMo 3** از اسناد علمی پردازش‌شده توسط خط لوله نوین **olmOCR** [۹۶] بهره می‌برد:

- **استخراج متن با مدل‌های بینایی-زبانی:** از میان ۲۳۸ میلیون سند PDF آکادمیک اولیه (تا تاریخ دسامبر ۲۰۲۴)، اسناد با ابزار olmOCR (نسخه‌های ۴۹.۱.۰ تا ۵۳.۱.۰) تبدیل به متن خطی و تمیز شدند و کتابخانه Poppler فقط در موارد اضطراری به کار رفت. این گام ۱۶۰ میلیون سند تحویل داد.
- **حذف داده‌های هویتی بافت‌آگاه (Context-Aware PII Removal):** برای جداسازی داده‌های هویتی خصوصی از موارد عمومی، یک خط لوله مدل‌محور دو سطحی طراحی شد: مدل Gemma-3-12B صفحه اول سند را برای یافتن هویت‌های حساس پایش می‌کرد و مدل Gemma-3-4B با تحلیل ۵۰۰۰ کاراکتر ابتدایی، ماهیت سند را بازشناسی می‌نمود. این تفکیک هوشمند منجر به پالایش ۹.۴٪ اسناد و صیانت از حریم شخصی شد.
- **پالایش نهایی و پروانه نشر:** حذف اسناد با تراکم بیش از ۳۰٪ جدول، متون دارای بیش از ۲۰٪ محتوای صرفاً عددی، تبدیل جداول Markdown به کدهای ساختاریافته HTML و فیلترینگ سخت‌گیرانه برای حفظ اسناد دارای پروانه‌های آزاد و تجاری، مجموعه‌ای شامل **۱۰۸ میلیون سند علمی (۹۷۲ میلیارد توکن)** به بار آورد.

صورت‌بندی ریاضی ترکیب داده‌ها و نمونه‌برداری آگاه از کیفیت (Olmix & Upsampling) انتخاب داده‌ها از میان ۴۸۰ سطل استخر وب و سایر منابع، از طریق حل یک مسئله بهینه‌سازی مقید ریاضی دو بخشی انجام پذیرفت:

۱. چارچوب ترکیب داده‌های مقید (Olmix Framework): با الهام از روش‌های ازدحامی [۶۰، ۹۷، ۹۸]، یک کلونی از مدل‌های پروکسی ۳۰ میلیون پارامتری با بودجه محاسباتی بهینه چینچیلا ($5\times$ معادل ۳ میلیارد توکن) روی توزیع‌های دیریکله نمونه‌برداری‌شده از حوزه‌ها آموزش دیده و عملکرد آن‌ها بر حسب معیار بیت بر بایت (Bits-per-Byte - BPB) در آزمون Base Easy Suite اندازه‌گیری شد. سپس یک مدل خطی تعمیم‌یافته (GLM) برای هر تسک برازش گردید:

$$\min_{\mathbf{w}} \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} g_t(\mathbf{w}) \quad \text{to subject} \quad \sum_i w_i = 1, \quad w_i \leq \frac{M \cdot X_i}{Z_{\text{total}}} \tag{۲۷}$$

که در آن $\mathbf{w}$ بردار وزن دامنه‌ها، $g_t(\cdot)$ مدل برآورد BPB تسک t، X_i تعداد کل توکن‌های در دسترس در دامنه i، $Z_{\text{total}} = 5.93\text{T}$ کل بودجه توکن و $M \in [4, 7]$ سقف مجاز فاکتور تکرار هر حوزه است. فرآیند در سه راند ترکیب شرطی (Conditional Mixing) تثبیت شد تا سهم وب به ۷۵٪ مقید شده و زبان‌های کدهای StackEdu [۹۹، ۱۰۰] و موضوعات مقالات علمی بهینه‌سازی شوند.

۲. نمونه‌برداری افزایشی بر مبنای توابع توانی-نمایی بریده‌شده (Quality-Aware Upsampling): به جای فیلترینگ ایستا (Flat Filtering) که اسناد را تنها به صورت دودویی حذف یا تک‌نسخه‌ای حفظ می‌کند، از خانواده توابع تحلیلی پیوسته و محدب زیر برای وزن‌دهی صدک‌های کیفی داده‌های وب استفاده شد:

$$f_{p,\lambda}(x) = \begin{cases} 0, & x < a \\ C(x-a)^p \cdot e^{\lambda(x-a)}, & x \geq a \end{cases} \tag{۲۸}$$

به‌طوری‌که پارامترهای $p \geq 0$، $\lambda \geq 0$ و ضریب مقیاس C از طریق حل عددی دستگاه معادلات مقید زیر استخراج می‌شوند:

$$\int_0^1 f_{p,\lambda}(x)\,dx = \frac{Z}{X}, \qquad \frac{1}{b}\int_{1-b}^1 f_{p,\lambda}(x)\,dx \leq M \tag{۲۹}$$

در این پیکربندی، آستانه $a = 0.40$ تعیین شد (دور ریختن ۴۰٪ داده‌های ضعیف)، Z/X نسبت توکن تخصیص‌یافته توسط Olmix، و سقف تکرار برای b درصد از مرغوب‌ترین داده‌های انتهایی برابر با $M = 7$ لحاظ گردید. در نهایت، فاکتور بازنمونه‌برداری هر سطل صدکی $[\,q^-,\,q^+\,]$ از رابطه زیر به دست آمد:

$$\text{Factor Upsampling}_q = \frac{1}{q^+ - q^-}\int_{q^-}^{q^+} f_{p,\lambda}(x)\,dx \tag{۳۰}$$

اعتبارسنجی تجربی و نتایج ابلیشن دادگان پیش‌آموزش صحت کارایی روش‌های فوق با آموزش مدل‌های ۱ میلیارد پارامتری روی ۱۰۰ میلیارد توکن در بنچمارک Base Easy Suite اعتبارسنجی شد (کاهش معیار BPB بیانگر بهبود کیفیت است). نتایج جدول ۲۳ برتری قاطع رویکرد تلفیقی پیشنهادی را به اثبات می‌رساند.

جدول ۲۳: نتایج مطالعات حذف (Ablation) در انتخاب استراتژی ترکیب داده و نمونه‌برداری کیفی

استراتژی پیکربندی داده	(BPB) QA Easy	(BPB) Math Easy	(BPB) Code Easy
توزیع طبیعی داده‌ها (Natural Distribution)	۷۰.۱۰۰	۸۷.۷۱	۲۰.۵۹
ترکیب چندهدفه متوازن Olmix	**۵۴.۹۹**	**۶۹.۶۱**	**۹۴.۴۸**
فیلترینگ ایستا ۵۰٪ برتر (تکرار ۱.۱ برابر)	۱۹.۱۰۴	۲۸.۸۶	۳۴.۹۴
فیلترینگ ایستا ۵٪ برتر (تکرار ۱.۱۱ برابر)	۵۳.۱۰۶	۳۱.۸۴	۰۲.۹۳
نمونه‌برداری کیفی پیوسته **OLMo 3**	**۹۵.۹۹**	**۰۴.۷۴**	**۸۸.۷۱**
تلفیق با میانگین هندسی (Geometric Mean)	۴۲.۱۰۰	۱۸.۷۸	۳۴.۸۱
تلفیق با تابع نمایی ساده ($Ce^{\lambda(x-a)}$)	۱۷.۱۰۰	۲۴.۷۸	۷۲.۷۸
تلفیق توانی-نمایی بریده‌شده (OLMo 3 Mix)	**۳۱.۹۹**	**۸۱.۷۵**	**۲۶.۷۸**

پیکربندی محاسباتی و هایپرپارامترهای پیش‌آموزش آموزش مدل‌های پایه بر بستر زیرساخت پرسرعت OLMo-core در قالب مخزن توزیع‌پذیر PyTorch FSDP2 اجرا گردید. مشخصات فنی این مرحله در جدول ۲۴ گزارش شده است. جهت ارزیابی پیوسته کیفیت مدل ۳۲ میلیارد پارامتری بدون نیاز به توقف‌های پرهزینه بازپخت، روش میانگین‌گیری دوره‌ای وزن‌ها از ۴ چک‌پوینت متوالی با فاصله ۱۰۰۰ گام [۱۰۱] به کار گرفته شد.

۲.۱.۱ پس‌آموزش (Post-Training)

- **DeepSeek-V4**

مقدمه و پارادایم نوین دادگان پس‌آموزش فاز پس‌آموزش (Post-Training) در سری مدل‌های **DeepSeek-V4**—شامل دو نگارش DeepSeek-V4-Flash (با ۲۸۴ میلیارد پارامتر کل و ۱۳ میلیارد پارامتر فعال) و DeepSeek-V4-Pro (با ۶.۱ تریلیون پارامتر کل و ۴۹ میلیارد پارامتر فعال)—شاهد یک دگرگونی بنیادین روش‌شناختی نسبت به نسل‌های پیشین بوده است. در این نسخه، رویکرد متداول یادگیری تقویتی ترکیبی (Mixed Reinforcement Learning) که با چالش رقابت گرادیان‌ها و تضعیف هم‌زمان مهارت‌های ناهمگن روبرو بود، کاملاً با یک پارادایم دومرحله‌ای جایگزین گردید:

جدول ۲۴: هایپرپارامترهای فاز پیش‌آموزش مدل‌های پایه **OLMo 3**

پارامتر تنظیماتی (Hyperparameter)	OLMo 3 Base-7B	OLMo 3 Base-32B
تعداد لایه‌ها (Layers)	۳۲	۶۴
بعد پنهان (d_{model})	۴۰۹۶	۵۱۲۰
سازوکار توجه (Attention Heads)	۳۲ سر Q / ۳۲ سر KV	۴۰ سر Q / ۸ سر KV (GQA)
طول پنجره زمینه (Sequence Length)	۸،۱۹۲ توکن	۸،۱۹۲ توکن
توجه پنجره لغزان (Sliding Window Attention)	پنجره ۴۰۹۶ در ۳/۴ لایه‌ها	پنجره ۴۰۹۶ در ۳/۴ لایه‌ها
کل توکن‌های آموزش فاز اول	۹۳.۵ تریلیون توکن	۷۰.۵ تریلیون توکن
برنامه نرخ یادگیری (LR Schedule)	کسینوسی تعدیل‌شده تا ۹T.۵	کسینوسی منقطع در ۷T.۵
اوج نرخ یادگیری (Peak Learning Rate)	3.0×10^{-4}	6.0×10^{-4}
حداقل نرخ یادگیری نهایی (Final LR)	3.0×10^{-5}	6.21×10^{-5}
اندازه دسته در سطح توکن (Batch Size)	۴،۱۹۴،۳۰۴ توکن	۸،۳۸۸،۶۰۸ توکن
دمای اوج آموزش ($\mathrm{LR}^2/\mathrm{bsz}$)	2.146×10^{-14}	4.292×10^{-14}
ضریب Z-Loss	10^{-5}	10^{-5}
سخت‌افزار پردازشی	۱۰۲۴ کارت NVIDIA H100	۱۰۲۴ کارت NVIDIA H100
توان عملیاتی به ازای هر پردازنده	۷،۷۰۰ توکن/ثانیه (MFU ۴۳%)	۱،۹۰۰ توکن/ثانیه (MFU ۴۱%)

۱. **پرورش مستقل متخصصان حوزه‌ای (Specialist Training):** بهینه‌سازی مدل‌های تخصصی مستقل در حوزه‌های هدف از طریق تلفیق تنظیم دقیق نظارت‌شده (SFT) و یادگیری تقویتی حوزه‌ای (Domain-Specific RL).

۲. **تجمیع یکپارچه از طریق تقطیر برخط خط‌مشی (On-Policy Distillation - OPD):** ادغام دانش و توزیع خروجی بیش از ۱۰ مدل معلم متخصص در یک مدل نهایی دانش‌آموز بر روی خطوط سیر تولیدشده توسط خود دانش‌آموز [۱۰۲].

مرحله اول: آموزش متخصصان حوزه‌ای و مهندسی داده‌های SFT و RL در فاز نخست، مدل پایه با بهره‌گیری از داده‌های حوزه‌ای دست‌چین‌شده با چگالی اطلاعاتی بالا در چهار محور کلیدی آموزش داده می‌شود:

- **داده‌های تنظیم دقیق نظارت‌شده (SFT Data):** شامل نمونه‌های غنی و گام‌به‌گام استدلال زنجیره تفکر (CoT) در ریاضیات، داده‌های تعاملی حل مسائل چندمرحله‌ای نرم‌افزاری و متون دقیق نگارشی.
- **یادگیری تقویتی با الگوریتم GRPO:** بهینه‌سازی خط‌مشی با الگوریتم بهینه‌سازی نسبی خط‌مشی گروهی (GRPO) [۱۰۳] صورت می‌پذیرد که بدون نیاز به مدل نقاد مستقل، مزیت نسبی خروجی‌ها را از میان یک گروه G تایی از پاسخ‌های هم‌پرومپت محاسبه می‌کند:

$$A_i = \frac{R_i - \mathrm{mean}(\{R_1, \ldots, R_G\})}{\mathrm{std}(\{R_1, \ldots, R_G\})} \tag{۳۱}$$

شرط‌گذاری داده‌ها بر پایه تلاش استنتاجی (Reasoning Effort) توزیع خروجی و طول استدلال مدل از طریق تعیین جریمه‌های متغیر طول (Length Penalties) و پنجره‌های بافت در فاز RL به سه حالت رفتاری متمایز کالیبره شده است (جدول ۲۵).

جدول ۲۵: مقایسه حالات سه‌گانه تلاش استنتاجی و الزامات ساختاری داده‌ها در **DeepSeek-V4**

حالت استدلال (Mode)	ویژگی‌های رفتاری داده	سناریوی کاربردی	قالب پاسخ (Response Format)
غیرتفکری (Non-think)	پاسخ‌های شهودی و سریع بر پایه عادات آماری	وظایف روزمره و تصمیم‌گیری کم‌ریسک	</think> summary
تفکر عمیق (Think High)	تحلیل منطقی و کندتر با تمرکز بر دقت بالا	حل مسئله پیچیده و برنامه‌ریزی	<think> thinking tokens </think> summary
حداکثر تفکر (Think Max)	کشش استدلال تا نهایت مرزهای تحلیلی مدل	کاوش در دشوارترین مسائل منطقی	<think> ... </think> summary + پرامپت سیستمی ویژه

در حالت Think Max، پرامپت با تزریق یک دستورالعمل الزام‌آور سیستمی هدایت می‌شود:

Injected Instruction: *Reasoning Effort: Absolute maximum with no shortcuts permitted. You MUST be very thorough in your thinking and comprehensively decompose the problem to resolve the root cause, rigorously stress-testing your logic against all potential paths, edge cases, and adversarial scenarios. Explicitly write out your entire deliberation process, documenting every intermediate step, considered alternative, and rejected hypothesis to ensure absolutely no assumption is left unchecked.*

مدل پاداش زایشی (GRM) بر پایه داده‌های هدایت‌شده با شیوه‌نامه در تکالیف فاقد آزمون‌گر عینی، مدل‌های سنتی پاداش اسکالر که مبتنی بر ترجیحات انسانی (RLHF) آموزش می‌دیدند کنار گذاشته شده‌اند:

- **داده‌های روباریک‌محور (Rubric-Guided Data):** برای هر دسته از وظایف باز، شیوه‌نامه‌های تفصیلی تدوین گردید.
- **یگانگی ارزیاب و عامل زایا:** شبکه عامل اصلی (Actor Network) خود هم‌زمان در نقش مدل پاداش زایشی (GRM) عمل می‌کند. بهینه‌سازی تقویتی مستقیم بر روی رفتار قضاوت‌گری GRM، قوه استدلال درونی مدل را در فرایند امتیازدهی ادغام کرده و اتکا به داده‌های برچسب‌خورده انسانی را به حداقل رسانده است.

پروتکل‌های ساختاری داده: ابزارها، تفکر متناوب و دستورات سریع مدیریت بافت متنی عظیم یک میلیون توکنی در محیط‌های تعاملی نیازمند قالب‌بندی‌های داده‌ای دقیق است:

- **شمای فراخوانی ابزار با نشانگر |DSML|:** برای پیشگیری از نقص‌های متداول در قالب JSON، ساختاری بر پایه نشانه‌گذاری XML و توکن ویژه |DSML| توسعه یافت:

```
<|DSML|tool_calls>
<|DSML|invoke name="$TOOL_NAME">
<|DSML|parameter name="$PARAM_NAME" string="true|false">$PARAM_VALUE</|DSML|parameter>
</|DSML|invoke>
</|DSML|tool_calls>
```

- **استدلال متناوب (Interleaved Thinking):** برخلاف رویکردهای پیشین، در سناریوهای ابزارمحور کلیه ردپاهای فکری (Thinking Traces) در خلال نوبت‌های متوالی کاربر حفظ می‌گردد تا زنجیره استدلال در افق‌های زمانی بلند گسسته نشود؛ در حالی که در گفتگوهای عمومی جهت حفظ ظرفیت بافت، تاریخچه تفکر نوبت‌های قبلی حذف می‌شود.
- **توکن‌های دستور سریع (Quick Instruction):** جهت حذف سربار مدل‌های کوچک کمکی و جلوگیری از بازپردازش پرهزینه پیش‌تکمیل (Prefill)، توکن‌های اختصاصی نظیر <|action|> (تشخیص نیاز به وب‌گردی)، <|title|> (ساخت عنوان)، <|query|> (تولید کلیدواژه جستجو)، <|authority|> (سنجش اعتبار مرجع)، <|domain|> (شناسایی دامنه) و <|read_url|> (تصمیم‌گیری خواندن آدرس وب) مستقیماً درون بافت پردازش می‌شوند.

مرحله دوم: تقطیر برخط خط‌مشی (On-Policy Distillation - OPD) به‌منظور تثبیت یکپارچه دانش N مدل معلم تخصصی ($\pi_{E_1}, \ldots, \pi_{E_N}$ با $N > 10$) در قالب مدل واحد π_θ، تابع هزینه تقطیر برخط بر مبنای دیورژانس کولبک-لایبلر معکوس بر روی نمونه‌های مشتق‌شده از توزیع خود دانش‌آموز بهینه‌سازی می‌شود:

$$\mathcal{L}_{\mathsf{OPD}}(\theta) = \sum_{i=1}^{N} w_i \cdot \mathbb{D}_{\mathsf{KL}}\left(\pi_\theta \parallel \pi_{E_i}\right) = \sum_{i=1}^{N} w_i \cdot \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(\cdot|x)}\left[\log \frac{\pi_\theta(y \mid x)}{\pi_{E_i}(y \mid x)}\right] \tag{۳۲}$$

که در آن w_i وزن تخصیص‌یافته به تخصص معلم i-ام است. برخلاف تقریب‌های تک‌توکنی مرسوم:

$$\mathsf{sg}\left[\log \frac{\pi_{E_i}(y_t \mid x, y_{<t})}{\pi_\theta(y_t \mid x, y_{<t})}\right] \tag{۳۳}$$

که نوسان شدید شیب ایجاد می‌کردند، در **DeepSeek-V4** تقطیر بر روی لاجیت‌های کل دایره واژگان ($|V| > 100$K) پیاده‌سازی شده است. همچنین جهت جلوگیری از اریب آماری طول توالی (Length Bias) در صورت بروز وقفه حین تولید بافت‌های میلیونی، سازوکار ثبت پیش‌نویس توکن‌محور (Token-Granular WAL) مستقر گردیده تا رمزگشایی دقیقاً از توکن متوقف‌شده از سر گرفته شود.

ارزیابی‌های تجربی روی بنچمارک‌های استاندارد (Standard Evaluations) نتایج ارزیابی مدل نهایی پس‌آموزش‌دیده DeepSeek-V4-Pro-Max در جدول ۲۶ و بررسی اثر حالات تفکر در جدول ۲۷ آورده شده است. ارقام گویای آن است که مدل در استدلال رقابتی به ریتینگ کدفرسز 3206 (رتبه ۲۳ انسانی در جهان) و دقت 93.5 در LiveCodeBench دست یافته و در آزمون حقیقت‌سنجی SimpleQA-Verified با امتیاز 57.9، رکورد مدل‌های متن‌باز را به میزان ۲۰ درصد مطلق جابجا کرده است.

جدول ۲۶: مقایسه عملکرد DeepSeek-V4-Pro-Max در برابر مدل‌های رقیب متن‌باز و تجاری

دسته ارزیابی	بنچمارک (متریک)	Opus-4.6-Max	GPT-5.4-xHigh	Gemini-3.1-Pro	K2.6-Thinking	GLM-5.1-Thinking	DS-V4-Pro-Max
دانش و استدلال	MMLU-Pro (EM)	۱.۸۹	۵.۸۷	**۰.۹۱**	۱.۸۷	۰.۸۶	۵.۸۷
	SimpleQA-Verified (Pass@1)	۲.۴۶	۳.۴۵	**۶.۷۵**	۹.۳۶	۱.۳۸	۹.۵۷
	Chinese-SimpleQA (Pass@1)	۴.۷۶	۸.۷۶	**۹.۸۵**	۹.۷۵	۰.۷۵	۴.۸۴
	GPQA Diamond (Pass@1)	۳.۹۱	۰.۹۳	**۳.۹۴**	۵.۹۰	۲.۸۶	۱.۹۰
	HLE (Pass@1)	۰.۴۰	۸.۳۹	**۴.۴۴**	۴.۳۶	۷.۳۴	۷.۳۷
	LiveCodeBench (Pass@1)	۸.۸۸	–	۷.۹۱	۶.۸۹	–	**۵.۹۳**
	Codeforces (Rating)	–	۳۱۶۸	۳۰۵۲	–	–	**۳۲۰۶**
	HMMT 2026 Feb (Pass@1)	۲.۹۶	**۷.۹۷**	۷.۹۴	۷.۹۲	۴.۸۹	۲.۹۵
	IMOAnswerBench (Pass@1)	۳.۷۵	**۴.۹۱**	۰.۸۱	۰.۸۶	۸.۸۳	۸.۸۹
	Apex Shortlist (Pass@1)	۹.۸۵	۱.۷۸	۱.۸۹	۵.۷۵	۴.۷۲	**۲.۹۰**
بافت طولانی	MRCR 1M (MMR)	**۹.۹۲**	–	۳.۷۶	–	–	۵.۸۳
	CorpusQA 1M (ACC)	**۷.۷۱**	–	۸.۵۳	–	–	۰.۶۲
عملکرد عاملی	Terminal Bench 2.0 (Acc)	۴.۶۵	**۱.۷۵**	۵.۶۸	۷.۶۶	۵.۶۳	۹.۶۷
	SWE-Verified (Resolved)	**۸.۸۰**	–	۶.۸۰	۲.۸۰	–	۶.۸۰
	SWE-Pro (Resolved)	۳.۵۷	۷.۵۷	۲.۵۴	**۶.۵۸**	۴.۵۸	۴.۵۵
	BrowseComp (Pass@1)	۷.۸۳	۷.۸۲	**۹.۸۵**	۲.۸۳	۳.۷۹	۴.۸۳
	MCPAtlas Public (Pass@1)	**۸.۷۳**	۲.۶۷	۲.۶۹	۶.۶۶	۸.۷۱	۶.۷۳
	Toolathlon (Pass@1)	۲.۴۷	**۶.۵۴**	۸.۴۸	۰.۵۰	۷.۴۰	۸.۵۱

جدول ۲۷: مقایسه تاثیر حالات تلاش تفکر بر نسخه‌های Flash و Pro مدل **DeepSeek-V4**

بنچمارک (متریک)	DeepSeek-V4-Flash			DeepSeek-V4-Pro		
	Non-Think	High	Max	Non-Think	High	Max
MMLU-Pro (EM)	۰.۸۳	۴.۸۶	۲.۸۶	۹.۸۲	۱.۸۷	۵.۸۷
SimpleQA-Verified (Pass@1)	۱.۲۳	۹.۲۸	۱.۳۴	۰.۴۵	۲.۴۶	۹.۵۷
Chinese-SimpleQA (Pass@1)	۵.۷۱	۲.۷۳	۹.۷۸	۸.۷۵	۷.۷۷	۴.۸۴
GPQA Diamond (Pass@1)	۲.۷۱	۴.۸۷	۱.۸۸	۹.۷۲	۱.۸۹	۱.۹۰
HLE (Pass@1)	۱.۸	۴.۲۹	۸.۳۴	۷.۷	۵.۳۴	۷.۳۷
LiveCodeBench (Pass@1-COT)	۲.۵۵	۴.۸۸	۶.۹۱	۸.۵۶	۸.۸۹	۵.۹۳
Codeforces (Rating)	–	۲۸۱۶	۳۰۵۲	۲۹۱۹	۳۲۰۶
HMMT 2026 Feb (Pass@1)	۸.۴۰	۹.۹۱	۸.۹۴	۷.۳۱	۰.۹۴	۲.۹۵
MRCR 1M (MMR)	۵.۳۷	۹.۷۶	۷.۷۸	۷.۴۴	۳.۸۳	۵.۸۳
CorpusQA 1M (ACC)	۵.۱۵	۳.۵۹	۵.۶۰	۶.۳۵	۵.۵۶	۰.۶۲
SWE-Verified (Resolved)	۷.۷۳	۶.۷۸	۰.۷۹	۶.۷۳	۴.۷۹	۶.۸۰

ارزیابی‌های کیفی بر روی دادگان کاربردی جهان واقعی به‌منظور پر کردن شکاف میان بنچمارک‌های ایستا و تجربه واقعی کاربران، مجموعه‌داده‌های داخلی زیر ارزیابی گردیدند:

- **نگارش عملکردی زبان چینی (۳،۱۷۰ نمونه):** طبق جدول ۲۸، مدل در برابر Gemini-3.1-Pro به نرخ برد کلی 62.65% دست یافت.

- **نگارش خلاقانه چینی (۲،۸۳۷ نمونه):** مطابق جدول ۲۹، مدل در کیفیت نویسندگی به نرخ برتری چشمگیر 77.48% در برابر رقیب رسید.
- **دستورات پیچیده و نگارش چندنوبته (۱۹۶ پرامپت چالش‌برانگیز):** طبق جدول ۳۰، رقابت فشرده‌ای میان DeepSeek-V4-Pro با نرخ 45.9% و Claude-Opus-4.5 با 52.0% ثبت شد.
- **ارزیابی جستجوی عاملی در برابر RAG:** آزمون بر روی ۸۶۹ پرامپت (جدول ۳۱) نشان داد که جستجوی عاملی با 61.7% برد، رویکرد RAG را پشت سر گذاشته است؛ در حالی که طبق ارزیابی ۹۵۶ پرامپت جستجو، مدل V4 با برتری 28.1% در برابر 10.4% نسبت به DeepSeek-V3.2 غلبه دارد.
- **تکالیف اداری-تخصصی سفیدپوشان (White-Collar Tasks):** ارزیابی انسانی ۳۰ سناریوی پیشرفته اداری در ۱۳ صنعت حاکی از نرخ عدم باخت 63% نسبت به Opus-4.6-Max و برتری در تکمیل تسک (98.32 در برابر 96.68) و غنای محتوا (83.32 در برابر 78.00) است.
- **عامل‌های مهندسی نرم‌افزار در محیط R&D:** طبق جدول ۳۲، ارزیابی روی ۳۰ تسک استانداردسازی‌شده از مسائل واقعی مهندسان داخلی نشان داد **DeepSeek-V4-Pro-Max** با نرخ موفقیت 67%، مدل Claude Sonnet 4.5 (47%) را قاطعانه پشت سر گذاشته و در مجاورت Opus 4.5 (70%) قرار گرفته است.

جدول ۲۸: خلاصه نتایج نگارش عملکردی چینی در برابر Gemini-3.1-Pro (۳،۱۷۰ نمونه)

دسته نوشتاری	تعداد (#)	برد DS	برد Gem	درصد برد DS	درصد برد Gem	درصد مساوی
متون اداری و تجاری (Business)	۱۳۴۹	۸۷۹	۴۳۶	۱۶%.۶۵	۳۲%.۳۲	۵۲%.۲
متون رسانه‌ای و تبلیغاتی (Media)	۶۶۶	۳۸۶	۲۵۶	۹۶%.۵۷	۴۴%.۳۸	۶۰%.۳
نگارش روزمره و ارتباطی (Everyday)	۳۹۰	۲۷۱	۱۰۱	۴۹%.۶۹	۹۰%.۲۵	۶۲%.۴
متون گفتاری و سخنرانی (Oral)	۳۱۹	۱۸۷	۱۲۰	۶۲%.۵۸	۶۲%.۳۷	۷۶%.۳
اسناد رسمی و نامه‌های دولتی (Official)	۲۳۰	۱۲۶	۹۷	۷۸%.۵۴	۱۷%.۴۲	۰۴%.۳
متون آکادمیک و ترویج علم (Academic)	۲۱۶	۱۳۷	۷۱	۴۳%.۶۳	۸۷%.۳۲	۷۰%.۳
مجموع کل	**۳۱۷۰**	**۱۹۸۶**	**۱۰۸۱**	**۶۵%.۶۲**	**۱۰%.۳۴**	**۲۵%.۳**

جدول ۲۹: ارزیابی نگارش خلاقانه چینی در برابر Gemini-3.1-Pro (۲،۸۳۷ نمونه)

محور ارزیابی	آمار شمارشی			درصد آماری		
	برد DS	برد Gem	مساوی	% برد DS	% برد Gem	% مساوی
پیروی از دستورالعمل (Instruction Following)	۱۷۰۳	۱۱۱۹	۱۵	۰۳%.۶۰	۴۴%.۳۹	۵۳%.۰
کیفیت ادبی و نویسندگی (Writing Quality)	۲۱۹۸	۶۳۴	۵	۴۸%.۷۷	۳۵%.۲۲	۱۸%.۰

جدول ۳۰: ارزیابی دستورات فوق‌پیچیده و نگارش چندنوبته در برابر Claude Opus 4.5

دسته‌بندی	تعداد (#)	برد DS	برد Opus	درصد DS	درصد Opus	درصد مساوی
پیروی از دستورات پیچیده (Complex Instructions)	۴۹	۲۳	۲۶	۹%.۴۶	۱%.۵۳	۰%.۰
نگارش چندنوبته (Multi-Turn Writing)	۱۴۷	۶۷	۷۶	۶%.۴۵	۷%.۵۱	۷%.۲
مجموع کل	**۱۹۶**	**۹۰**	**۱۰۲**	**۹%.۴۵**	**۰%.۵۲**	**۰%.۲**

- **GLM-5**

مقدمه و پارادایم داده‌های پس‌آموزش فاز پس‌آموزش (Post-Training) در مدل **GLM-5** با هدف گذار ساختاری از پارادایم «کدنویسی حسی» (Vibe Coding)—که مبتنی بر پرامپت‌نویسی انسانی است—به سمت «مهندسی عامل‌محور» (Agentic Engineering) بازطراحی شده است؛ جایی که مدل به عنوان یک عامل خودکار، وظایف برنامه‌ریزی، پیاده‌سازی، اجرا و خوداصلاحی را در افق‌های زمانی طولانی بر

جدول ۳۱: سنجش کارایی جستجوی عاملی در مقایسه با RAG سنتی در DeepSeek-V4-Pro

سطح مسئله	نوع پرسش	تعداد	برد عامل	برد RAG	% برد عامل	% برد RAG	% مساوی
آسان (Easy)	عینی (Objective)	۱۹۶	۱۱۰	۴۳	۱%.۵۶	۹%.۲۱	۹%.۲۱
	تحلیلی (Subjective)	۳۲۱	۱۹۸	۵۶	۷%.۶۱	۴%.۱۷	۹%.۲۰
دشوار (Hard)	عینی (Objective)	۱۶۸	۱۰۲	۳۳	۷%.۶۰	۶%.۱۹	۶%.۱۹
	تحلیلی (Subjective)	۱۸۴	۱۲۶	۲۷	۵%.۶۸	۷%.۱۴	۸%.۱۶
جمع کل	–	**۸۶۹**	**۵۳۶**	**۱۵۹**	**۷%.۶۱**	**۳%.۱۸**	**۰%.۲۰**

جدول ۳۲: مقایسه نرخ موفقیت در حل وظایف مهندسی نرم‌افزار واقعی (R&D Benchmark)

مدل	Haiku 4.5	Sonnet 4.5	DS-V4-Pro-Max	Opus 4.5	Opus 4.5 (Think)	Opus 4.6 (Think)
نرخ حل (Pass Rate %)	۱۳%	۴۷%	**۶۷%**	۷۰%	۷۳%	۸۰%

عهده می‌گیرد. برای این منظور، خط لوله داده‌های پس‌آموزش با بسط طول بافت تا سقف **۲۰۲،۷۵۲ توکن** در قالب یک زنجیره چندمرحله‌ای شامل تنظیم دقیق بانظارت (SFT)، یادگیری تقویتی استدلال (Reasoning RL)، یادگیری تقویتی عامل‌محور (Agentic RL)، یادگیری تقویتی عمومی (General RL) و تقطیر هم‌خط بین‌مرحله‌ای (On-Policy Cross-Stage Distillation) سازمان‌دهی شده است.

ترکیب‌بندی و پالایش پیکره داده‌های تنظیم دقیق بانظارت (SFT Data) پیکره داده‌های SFT در **GLM-5** نسبت به نسل‌های پیشین به شکل چشمگیری در حوزه‌های کدنویسی و تعاملات عاملی گسترش یافته و در سه محور ساختار یافته است:

- **گفت‌وگوی عمومی (General Chat):** شامل داده‌های پرسش و پاسخ، نگارش، ترجمه، تعاملات طولانی و بازی نقش (Role-playing) چندزبانه. داده‌های بازی نقش بر پایه پنج بعد کیفی (پیروی از دستور، بیان‌گری زبانی، خلاقیت، انسجام منطقی و پایداری در مکالمات طولانی) توسط ارزیاب‌های خودکار و متخصصان انسانی پالایش شدند. سبک پاسخ‌ها نیز به سوی ساختاری منطقی‌تر و موجزتر بهینه‌سازی شد.
- **داده‌های استدلال تحلیلی (Reasoning Data):** در حوزه استدلال منطقی، داده‌ها از مسائل قابل راستی‌آزمایی عینی با روش نمونه‌برداری حذفی (Rejection Sampling) استخراج شدند. در مسائل ریاضی و علوم، فیلتر مبتنی بر دشواری (Difficulty-based Filtering) اعمال گردید تا تنها مسائلی حفظ شوند که مدل **GLM-4.7** توانایی حل آن‌ها را نداشته است.
- **داده‌های کدنویسی و عاملی (Coding & Agent Data):** شامل کدهای فرانت‌اند و بک‌اند، فراخوانی ابزارها و ردپای عامل‌های حل مسأله در محیط‌های واقعی. داده‌ها با الگوریتم‌های یادگیری تقویتی خبره و نمونه‌برداری حذفی استخراج شدند. بخش‌های اشتباه در مسیر اجرای عامل حذف نشدند، بلکه در تابع زیان آموزش ماسک‌گذاری شدند (Masked out):

$$\mathcal{L}_{\text{SFT}}(\theta) = -\sum_{t \in \mathcal{T}_{\text{valid}}} \log \pi_\theta(y_t \mid x, y_{<t}) \tag{۳۴}$$

این رویکرد به مدل می‌آموزد که الگوهای تصحیح خطا (Error Correction) را بدون بازتولید یا تقویت اقدامات غلط فراگیرد.

قالب‌بندی داده‌ها و حالات سه‌گانه تفکر (Thinking Modes) قالب پرامپت‌ها در داده‌های پس‌آموزش به‌گونه‌ای مهندسی شده است که مدل را به سه الگوی رفتاری ساختاریافته مجهز می‌سازد:

- **تفکر درهم‌تنیده (Interleaved Thinking):** تعبیه بلوک‌های استدلال پیش از هر پاسخ متنی و پیش از هر فراخوانی ابزار (Tool Call) جهت بهبود دقت عملیاتی.

- **تفکر حفظ‌شده (Preserved Thinking):** در تعاملات چندنوبته مهندسی نرم‌افزار، زنجیره تفکر نوبت‌های پیشین در زمینه ذخیره می‌شود تا از استدلال مجدد از نقطه صفر و ایجاد تناقض‌های منطقی در وظایف طولانی جلوگیری شود.
- **تفکر در سطح نوبت (Turn-level Thinking):** قابلیت تنظیم فعال/غیرفعال بودن تفکر در هر نوبت بر اساس میزان پیچیدگی درخواست به منظور کاهش تأخیر زمانی.

داده‌ها و فرمولاسیون یادگیری تقویتی استدلال (Reasoning RL) در مرحله استدلال تحلیلی، داده‌ها در چهار حوزه متعادل توزیع شده‌اند:

- **ریاضیات و علوم:** گردآوری از دادگان متن‌باز معتبر نظیر Nemotron-Math [۱۰۴] و راهکار برنده رقابت AIMO-2 با پایگاه داده OpenMathReasoning [۱۰۵]. مسائل به گونه‌ای فیلتر شدند که برای **GLM-4.7** غیرقابل حل بوده، اما معلمان قدرتمند نظیر GPT-5.2 (xhigh) [۱۰۶] و Gemini 3 Pro Preview [۱۰۷] توانایی حل گام‌به‌گام آن‌ها را داشته باشند.
- **کدنویسی رقابتی و علمی:** مسائل مسابقات برنامه‌نویسی از پلتفرم Codeforces، مجموعه‌داده‌های TACO [۱۰۸] و SYNTHETIC-2-RL [۱۰۹].
- **استدلال یکپارچه با ابزار (Tool-Integrated Reasoning - TIR):** نمونه مسائل ریاضی و مهندسی که مستلزم فراخوانی مفسر پایتون و ابزارهای محاسباتی بیرونی هستند.

آموزش بر پایه نسخه اصلاح‌شده الگوریتم GRPO [۲۴] و تکنیک IcePop [۱۱۰] برای حذف واگرایی میان سیاست استنتاج و سیاست گرادیان صورت گرفته است:

$$\mathcal{L}(\theta) = -\mathbb{E}_{x\sim\mathcal{D},\{y_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}^{\text{infer}}}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\text{pop}(\rho_{i,t},1/\beta,\beta)\right.$$
$$\left.\cdot\min\left(r_{i,t}\hat{A}_{i,t},\ \text{clip}(r_{i,t},1-\epsilon_{\text{low}},1+\epsilon_{\text{high}})\hat{A}_{i,t}\right)\right] \tag{۳۵}$$

که در آن نسبت عدم‌انطباق استنتاج-آموزش و عملگر سرکوب خطای pop عبارتند از:

$$\rho_{i,t} = \frac{\pi_{\theta_{\text{old}}}^{\text{train}}(y_{i,t}\mid x,y_{i,<t})}{\pi_{\theta_{\text{old}}}^{\text{infer}}(y_{i,t}\mid x,y_{i,<t})},\quad \text{pop}(\rho,1/\beta,\beta) = \begin{cases}\rho, & \text{if} 1/\beta\le\rho\le\beta\\ 0, & \text{otherwise}\end{cases} \tag{۳۶}$$

آموزش این بخش با اندازه گروه $G = 32$، اندازه دسته ۳۲ و هایپرپارامترهای $\beta = 2$، $\epsilon_{\text{low}} = 0.2$ و $\epsilon_{\text{high}} = 0.28$ انجام شده است.

محیط‌های اجرایی و ساخت دادگان عامل‌محور (Agentic RL Environments) برای آموزش مدل در نقش عامل مهندسی، داده‌ها از متن‌های ایستا فراتر رفته و در بستر بیش از ۱۰ هزار محیط اجرایی ابطال‌پذیر تعاملی تولید شدند:

- **محیط‌های مهندسی نرم‌افزار (SWE Environments):** با اقتباس از چارچوب RepoLaunch [۱۱۱]، بیش از **۱۰،۰۰۰ محیط اجرایی** بر پایه مخازن گیت‌هاب در ۹ زبان برنامه‌نویسی (Python، Java، Go، C، C++، JavaScript، TypeScript، PHP و Ruby) ایجاد شد. آزمون‌های شکست-به-موفقیت (Fail-to-Pass - F2P) و پایداری (Pass-to-Pass - P2P) توسط لاگ‌خوان‌های مبتنی بر مدل زبانی استخراج شدند.
- **محیط‌های ترمینال (Terminal Environments):** تولید داده‌ها در ساختار استاندارد کانتینری Harbor [۱۱۲] انجام گرفت. این فرآیند از دو مسیر انجام شد: ۱) سنتز سه‌مرحله‌ای از وظایف بذر مهندسی نرم‌افزار با دقت ساخت داکر بیش از ۹۰٪، ۲) استخراج خودکار و حلقه-بسته از صفحات وب فنی که در آن عامل با اجرای اسکریپت‌های اعتبارسنجی Harbor به صورت خودکار وظیفه تولیدی را اصلاح و نهایی می‌کرد.

- **وظایف جستجوی عمیق چندمرحله‌ای (Deep-Search Tasks):** پایگاه داده‌ای از گراف دانش وب (WKG) بر پایه ۲ میلیون صفحه با چگالی اطلاعاتی بالا ساخته شد. سوالات از روی زنجیره‌های چندموجودیتی استخراج و از یک فیلتر سه‌مرحله‌ای عبور داده شدند: ۱) حذف سوالات قابل حل بدون ابزار (در ۱ تلاش از ۸ اجرا)، ۲) حذف سوالات قابل حل با جستجوی تک‌مرحله‌ای، ۳) اعتبارسنجی دوطرفه تطابق شواهد وب با پاسخ مرجع توسط عامل داور.
- **محیط تولید اسلاید متنی-بصری (Slide Generation):** مدلسازی داده‌ها بر پایه ساختار HTML با پاداش سه‌سطحی: سطح ۱ (ویژگی‌های نشانه‌گذاری استایل، رنگ و حذف تصاویر توهمی)، سطح ۲ (ویژگی‌های رندرینگ زمان اجرا در DOM نظیر ابعاد و جلوگیری از هک پاداش با قطع نامتقارن متن)، و سطح ۳ (ارزیابی بصری توزیع فضاهای خالی). در این فرایند انطباق ساختاری با نسبت ابعاد ۱۶:۹ از ٪۴۰ به **٪۹۲** افزایش یافت.

پایدارسازی آماری داده‌ها در یادگیری تقویتی ناهمگام (Asynchronous Stability) تولید برون‌خط و هم‌زمان مسیرهای اجرایی عامل‌ها نیازمند کنترل آماری عدم‌انطباق سیاست‌ها بود:

- **دروازه توکن‌به‌توکن (TITO Gateway):** ثبت مستقیم شناسه‌های توکن و فراداده‌ها در خروجی موتور استنتاج به منظور رفع خطاهای ناشی از توکنایزیشن مجدد در سمت آموزش.
- **نمونه‌برداری اهمیتی دوطرفه مستقیم (Direct Double-sided Importance Sampling):** کنترل بازه گرادیان در محدوده $[1-\epsilon_\ell, 1+\epsilon_h]$ بدون نیاز به فراخوانی پرهزینه سیاست قدیمی:

$$\mathcal{L}(\theta) = \mathbb{E}_t\left[f(r_t(\theta), \epsilon_\ell, \epsilon_h)\hat{A}_t \log \pi_\theta(a_t \mid s_t)\right] \tag{۳۷}$$

$$r_t(\theta) = \exp\left(\log \pi_\theta(a_t \mid s_t) - \log \pi_{\text{rollout}}(a_t \mid s_t)\right) \tag{۳۸}$$

$$f(x; \epsilon_\ell, \epsilon_h) = \begin{cases} x, & \text{if} 1-\epsilon_\ell < x < 1+\epsilon_h \\ 0, & \text{otherwise} \end{cases} \tag{۳۹}$$

- **پالایش داده‌های بیات و خطاهای محیطی:** حذف مسیرهایی که اختلاف نسخه مدل تولیدکننده آن‌ها با نسخه جاری آموزش بیش از حد آستانه بود ($w' - w_0 > \tau$). همچنین حذف داده‌های حاصل از فروپاشی کانتینر داکر و تکمیل داده‌های گروهی (Padding) در صورت وجود حداقل نیمی از نمونه‌های معتبر.

یادگیری تقویتی عمومی و تقطیر هم‌خط بین‌مرحله‌ای در فاز یادگیری عمومی، اهداف بهینه‌سازی در سه بعد صحت بنیادین (کاهش توهم و خطاهای ساختاری)، هوش هیجانی (همدلی و گفت‌وگوی روان) و کیفیت تخصصی وظایف تنظیم شدند. برای نظارت بر این ابعاد، سیستمی ترکیبی از پاداش‌های قانون‌محور، مدل‌های پاداش پیامد (ORMs) و مدل‌های پاداش زایندگی (GRMs) به کار رفت. افزون بر این، به منظور ممانعت از تولید خروجی‌های بیش‌ازحد کلیشه‌ای و ماشینی، پاسخ‌های تدوین‌شده توسط متخصصان برجسته انسانی (Human-authored Responses) به عنوان لنگرهای کیفی سبک وارد جریان آموزش شدند.

در نهایت، برای رفع افت تدریجی توانمندی‌های پیشین (Catastrophic Forgetting)، تقطیر هم‌خط بین‌مرحله‌ای بر روی توزیع داده‌های مراحل استدلال و عمومی با اندازه دسته ۱۰۲۴ و اندازه گروه $G = 1$ اجرا شد:

$$\hat{A}_{i,t} = \text{sg}\left[\log \frac{\pi^{\text{infer}}_{\theta_{\text{teacher}}}(y_{i,t} \mid x, y_{i,<t})}{\pi^{\text{train}}_{\theta}(y_{i,t} \mid x, y_{i,<t})}\right] \tag{۴۰}$$

جدول ۳۳: توزیع سناریوهای کاربردی فرانت‌اند و واحدهای اعتبارسنجی در CC-Bench-V2

دسته‌بندی	شرح سناریو	تعداد وظایف (# Tasks)	تعداد چک‌لیست‌ها (# Checkitems)
سامانه‌های سازمانی (Business Systems)	مدیریت فرآیندها و داده‌های تجاری	۴۲	۱۶۷
بازی‌های تعاملی وب (Web Games)	تعاملات گرافیکی و برنامه‌های سرگرمی	۴۰	۱۶۳
رندرینگ برداری (SVG/Canvas)	ترسیمات پویا و مصورسازی‌های تعاملی	۳۲	۱۶۶
ابزارهای خلاقانه (Creative Tools)	ابزارهای ویرایش و تولید محتوای برخط	۲۸	۱۶۰
صفحات معرفی (Showcase Pages)	بازنمایی اطلاعات و طراحی صفحات فرود	۲۷	۱۱۵
فرم‌ها و جداول (Forms & Tables)	ساختارهای ورود داده و اعتبارسنجی ورودی	۲۶	۹۳
مصورسازی داده (Data Visualization)	نمودارهای تحلیلی و گرافیکی وب	۲۵	۸۵
مجموع کل وظایف فرانت‌اند	پوشش کامل پشته‌های HTML/JS (۱۱۳)، React (۵۸) و Vue (۴۹)	**۲۲۰**	**۹۶۹**

جدول ۳۴: ارزیابی عملکرد مدل در بخش‌های فرانت‌اند، بک‌اند و افق طولانی بنچمارک CC-Bench-V2

حوزه مهندسی	پشته یا نوع وظیفه	متریک ارزیابی	GLM-5	GLM-4.7	Claude Opus 4.5
فرانت‌اند (Frontend)	HTML/CSS/JS	(%) ISR / CSR	**۳.۷۶ / ۹.۳۸**	۹.۶۴ / ۴.۳۵	۲.۸۲ / ۲.۵۲
	React	(%) ISR / CSR	**۰.۷۱ / ۶.۳۴**	۴.۴۹ / ۲.۱۷	۷.۷۰ / ۷.۳۹
	Vue	(%) ISR / CSR	**۱.۷۷ / ۷.۳۲**	۸.۵۳ / ۵.۲۴	۳.۷۴ / ۹.۴۶
صحت ساخت (Build)	React / Vue / Svelte / Next	(%) BSR	**۰.۹۵ / ۱۰۰ / ۱۰۰ / ۱۰۰**	۰.۶۵ / ۰.۷۰ / ۰.۶۰ / ۰.۷۰	۰.۹۵ / ۱۰۰ / ۰.۹۰ / ۰.۸۰
بک‌اند (Backend)	مهندسی نرم‌افزار (۸۵ تسک در ۶ زبان)	(%) Pass@1	**۸.۲۵**	۶.۱۹	۹.۲۶
افق طولانی (Long-horizon)	ناوبری مخازن بزرگ (Repo Exploration)	(%) Pass@1	**۶.۶۵**	۸.۴۷	۵.۶۴
	وظایف زنجیره‌ای متوالی (Chained Tasks)	(%) Pass@1	**۳.۵۲**	۰.۴۳	۶.۶۱

ارزیابی‌های تجربی روی مجموعه‌داده‌های عامل‌محور و دنیای واقعی به منظور ارزیابی کارایی داده‌های پس‌آموزش در شرایط واقعی مهندسی، بنچمارک داخلی جامع CC-Bench-V2 طراحی گردید. آمار ساختار داده‌ها و نتایج عملکردی مدل در جدول‌های زیر خلاصه شده است.

همچنین مدل **GLM-5** بر روی مجموعه‌داده‌های عملیاتی ارزیابی گردید که دستاوردهای زیر حاصل شد:

- **ترجمه ماشینی واقعی (ZMultiTransBench):** شامل ۱،۲۲۰ نمونه از سناریوهای متداول کاربری در ۷ جفت‌زبان از مبدا چینی به اسپانیایی (۳۰۰)، روسی (۲۵۰)، فرانسوی (۲۲۰)، کره‌ای (۲۰۰)، ژاپنی (۱۵۰)، عربی (۵۰) و آلمانی (۵۰). امتیاز الو مدل از ۱۰۱۶ در **GLM-4.7** به **۱۰۵۰** در **GLM-5** افزایش یافت.
- **ترجمه زبانی دشوار (MENT-SNS [۱۱۳]):** ۷۵۳ جفت‌جمله انگلیسی-چینی دارای کنایه و استعاره؛ امتیاز الو از ۹۹۳ به **۱۰۱۳** ارتقا یافت.
- **گفت‌وگوی چندزبانه (ZMultiDialBench):** ۱۴۱ سناریوی چندزبانه ارزیابی‌شده توسط متخصصان؛ نمره میانگین از ۴۴.۸ به **۵۷.۸** بهبود یافت.
- **پیروی از دستورات پیچیده (IF-Badcase):** ۴۵۰ نمونه پرامپت با محدودیت‌های چندگانه استخراج‌شده از محیط استقرار؛ نرخ موفقیت از ۵۷.۸٪ به **۲۷.۸۳٪** رسید.
- **فراخوانی ابزار (ToolCall-Badcase):** ۲۰۰ نمونه استخراج‌شده از موارد شکست کاربران در محیط عملیاتی؛ دقت تطابق ساختار و آرگومان‌ها با جهشی خیره‌کننده از ۸۰.۶٪ به **۹۸.۹۵٪** رسید.

- **Kimi K3**

مقدمه و ساختار کلان دادگان پس‌آموزش (Post-Training Paradigm) فاز پس‌آموزش در مدل زبانی چندوجهی **Kimi K3** [۱۳] (با ۸.۲ تریلیون پارامتر کل و ۱۰۴ میلیارد پارامتر فعال در هر توکن) بر پایه یک پارادایم نوین برای مقیاس‌پذیری محاسبات زمان آزمون (Test-Time Scaling) در پنجره بافت فوق‌طولانی **یک میلیون توکن** (1M-token context) طراحی شده است. این فرآیند بر یک زنجیره سه‌مرحله‌ای یکپارچه استوار است:

- **تنظیم دقیق تحت نظارت (SFT):** ایجاد یک سیاست راه‌انداز سرد (Cold-Start Policy) با کیفیت فوق‌العاده بالا جهت تثبیت رفتارهای تعاملی و عاملی.
- **یادگیری تقویتی چندسطحی (Multi-Effort RL):** آموزش متخصصان حوزه‌ای در ۳ دامنه کلان و ۳ سطح تلاش استدلالی (Reasoning Effort) شامل $\{\text{low}, \text{high}, \text{max}\}$ که به پیدایش ۹ مدل متخصص مجزا منجر می‌شود.

- **تقطیر هم‌سیاست چندمعلمه (Multi-Teacher On-Policy Distillation - MOPD):** تجمیع و ادغام قابلیت‌های ۹ متخصص حوزه‌ای درون یک مدل واحد سراسری.

سریال‌سازی و ساختار بازنمایی داده‌ها: قالب گفتگوی XTML جهت استانداردسازی تعاملات پیچیده عاملی، داده‌ها با قالب نشانه‌گذاری اختصاصی XTML (eXtensible Token Markup Language) فرمت‌بندی شده‌اند. این ساختار با جایگزینی نشانه‌های متداول XML با نشانه‌های رزرو شده ساختاری شامل `[open]`، `[sep]`، `[close]` و `[end_of_msg]`، ابهام ناشی از توکنایزیشن در مرز المان‌ها را از بین می‌برد و بار مالیات هم‌ترازی (Alignment Tax) را به حداقل می‌رساند:

- **تفکیک زون‌های داده (Zoning & Caching):** پیام‌های گزینه‌ای سراسری (Global Options) نظیر اعلان ابزارها (`tool-declare`) و تعیین بودجه تلاش استدلالی (`thinking-effort`) پیش از تمامی پیام‌های ورودی قرار می‌گیرند تا ساختار حافظه نهان کلید-مقدار (KV Cache) ابطال نشود. در مقابل، گزینه‌های گذرا (One-Shot) مانند `tool_choice` و `response_format` در انتهای پیام‌ها الصاق می‌شوند.
- **معماری سه‌کاناله دستیار (Channels):** پاسخ مدل دستیار در سه کانال مجزا سازمان‌دهی می‌شود:
  ۱. کانال استدلال و تفکر خام (`[open]think[sep]`).
  ۲. کانال پاسخ نهایی کاربر (`[open]response[sep]`).
  ۳. کانال فراخوانی ساختاریافته ابزارها (`[open]tools[sep]`).

  در این قالب، سازوکار **تفکر حفظ‌شده** (Preserved Thinking) اعمال می‌گردد؛ به‌طوری‌که حتی در صورت خالی بودن کانال استدلال، ساختار آن در تاریخچه حفظ می‌شود.
- **فراخوانی ابزار با آرگومان‌های تایپ‌شده (Typed Arguments):** هر فراخوانی حامل صفت `tool` و شناسه `index` برای تطبیق دقیق نتایج فراخوانی‌های موازی است. آرگومان‌های متنی و کدها به صورت متن خام تزریق شده و از ساختارهای پیچیده و فراداده‌شده JSON اجتناب می‌گردد. همچنین ماسک‌گذاری زیان (Loss Masking) بر روی ورودی‌های پشتیبان JSON اعمال می‌شود.

دادگان پیش‌تنظیم تحت نظارت (Supervised Fine-Tuning - SFT) مجموعه‌داده SFT به منظور ایجاد پایه رفتاری قدرتمند برای فاز یادگیری تقویتی گردآوری شده است:

- **سنتز مسیرهای عاملی (Trajectory Synthesis):** داده‌های عاملی با استفاده از مدل‌های تخصصی حوزه‌ای نسل‌های پیشین **Kimi K2** [۱۴] و **Kimi K2.5** [۱۱۴] شبیه‌سازی و تولید شده‌اند.
- **پالایش و اعتبارسنجی چندمرحله‌ای:** تمامی خط سیرهای تولیدشده از فیلترهای چندمرحله‌ای صحت‌سنجی خودکار و حاشیه‌نویسی انسانی در حلقه (Human-in-the-loop) عبور داده شدند تا هرگونه انحراف و توهم رفتاری پالایش گردد.
- **انطباق داده‌ها با آموزش آگاه از کوانتایزیشن (QAT):** داده‌های ورودی همگام با استراتژی کوانتایزیشن مدل در فاز استقرار تنظیم شده‌اند؛ وزن‌های متخصصان MoE در فرمت MXFP4 و فعال‌سازها در فرمت MXFP8 [۱۱۵] شبیه‌سازی می‌شوند تا مدل در برابر خطای دقت پایین مقاوم گردد.

سنتز داده‌های یادگیری تقویتی با هدایت گراف دانش (KG-Guided Task Synthesis) جهت پوشش جامع حوزه‌های دانشی با عمق مفهومی بالا و اجتناب از داده‌های سطحی، از یک **گراف جهت‌دار بدون دور (DAG)** خودتوسعه‌دهنده استفاده شده است:

- **کاوش بازگشتی عامل‌ها (Agentic Graph Construction):** فرآیند با گره‌های هسته در حوزه‌های اصلی (شیمی، فیزیک، ریاضی، علوم کامپیوتر، برنامه‌نویسی و علوم زیستی) آغاز می‌شود. عامل‌ها با وب‌گردی مستقل مفاهیم مرتبط را کشف کرده و یال‌های سلسله‌مراتبی از مفاهیم عام به اتمیک رسم می‌کنند تا یک گراف عمیق شکل گیرد.

- **بازیابی چندسطحی و سنتز وظایف:** با نمونه‌برداری از گره‌های گراف در سطوح دانه‌بندی گوناگون و ادغام با مفاهیم والد، کوئری‌های پیچیده‌ای استخراج شده و با بازیابی مقالات علمی، مخازن کد و اسناد واقعی وب، وظایف استدلالی چالش‌برانگیز تولید می‌شوند.

محیط جعبه‌سفید یکپارچه و تنوع چارچوب‌ها (Unified White-Box RL Environment) برای مهار بیش‌برازش مدل به یک قالب یا سیستم پرامپت ثابت، یک محیط چندماژوله جعبه‌سفید طراحی شده که می‌تواند رفتار چارچوب‌های نام‌آشنایی نظیر Kimi Code [۱۱۶]، Claude Code [۱۱۷]، Codex [۱۱۸]، OpenClaw [۱۱۹] و Hermes [۱۲۰] را شبیه‌سازی کند. ترکیب پویای ابزارها، مهارت‌ها و حافظه‌ها، استواری مدل را در مواجهه با توزیع‌های مختلف تضمین می‌نماید.

حوزه‌های تخصصی وظایف و داده‌های یادگیری تقویتی داده‌های تقویتی در بسترهای تعاملی و ابزارمحور با ویژگی‌های زیر جمع‌آوری و اعمال شده‌اند:

- **بهینه‌سازی کرنل‌های محاسباتی GPU (Kernel Tasks):** داده‌ها از مخازن سطح‌بالای گیت‌هاب نظیر Flash Linear Attention [۱۲۱] استخراج شده و شامل پیاده‌سازی‌های CUDA، Triton [۱۲۲]، CuTe DSL، ThunderKittens [۱۲۳] و TileLang [۱۲۴] در فرمت‌های BF16، FP8 و FP4 هستند. تابع پاداش بر مبنای تطابق دقیق خروجی با نسخه مرجع PyTorch (پاداش صفر در صورت خطای عددی) و نزدیکی به سقف مدل سقف‌شیشه‌ای سخت‌افزار (Hardware Roofline) بین 0.5 تا 1.0 تنظیم شده و سیستم‌های ضدهک پاداش، ترفندهایی مانند CUDA Graph Replay را جریمه می‌کنند.
- **دستیار شخصی در محیط‌های پایدار (Personal Assistant Workflows):** با پیاده‌سازی ماک‌های کامل نرم‌افزارهای Gmail، Notion، Slack و Canvas، سناریوهای شغلی چندروزه با ده‌ها رویداد درهم‌تنیده شبیه‌سازی شده‌اند که هر مسیر اجرایی می‌تواند شامل **هزاران فراخوانی ابزار** و **میلیون‌ها توکن زمینه** باشد.
- **وظایف اجرای خودمختار (Autonomous Execution Tasks - AET):** مدل‌ها با پارادایم تأیید در حلقه (Verify-in-the-Loop) هدایت می‌شوند. پاداش صرفاً بر اساس بررسی وضعیت نهایی محیط توسط تأییدگرهای مستقل اعطا می‌شود (نه گزارش متنی خود مدل). نمونه این وظایف، مهندسی معکوس و بازسازی سامانه‌های نرم‌افزاری جعبه‌سیاه از طریق ارسال کوئری به اوراکل است.
- **توسعه وب و صحنه‌های تعاملی (Web Development):** تولید سایت‌های کامل، انیمیشن‌های 3D/WebGL، رابط‌های کاربردی و کدهای SVG. پاداش شامل آزمون‌های عملکردی قطعی و قضاوت مدل پاداش در بررسی تعاملی و بصری صفحات رندرشده است.
- **استدلال بصری تأییدپذیر (Verifiable Visual Reasoning):** مدل به مفسر پایتون در سندباکس متصل است و می‌تواند با نوشتن کد، تصاویر مسائل علمی و نمودارها را برش دهد، بزرگ‌نمایی کند و تبدیل‌های هندسی روی آن اعمال نماید تا با دریافت تصاویر حاصله به عنوان مشاهدات جدید، پاسخ را گام‌به‌گام استخراج کند.

فرمول‌بندی‌های ریاضی بودجه، مهار اطناب و تقطیر داده‌ها جهت مدیریت طول توکن‌ها و همگرایی پایدار داده‌ها، معادلات زیر بر فرآیند آموزش حاکم شده‌اند:

- **کنترل بودجه استدلال (Reasoning Effort):** به هر مسئله x یک بودجه توکن اولیه $b_0(x)$ تخصیص یافته و پاداش در صورت فراتر رفتن از حد آستانه با مقدار -1 جریمه می‌شود:

$$R_{\text{effort}}(x,y) = \begin{cases} R_{\text{task}}(x,y), & \text{if} T(y) \le \tau \cdot b_0(x) \\ -1, & \text{if} T(y) > \tau \cdot b_0(x) \end{cases} \tag{۴۱}$$

که در آن $T(y)$ در وظایف استدلالی تعداد توکن‌های تفکر و در وظایف عاملی کل توکن‌های خروجی است و ضریب τ در طول برنامه درسی کاهش می‌یابد.

- **مهار اطناب در مدل پاداش عاملی (GRM):** اگر طول متن خروجی کاندیدا از $\sigma \cdot \ell_0$ تجاوز کند (که ℓ_0 طول پاسخ اولیه مدل پایه است)، کاندیدا به صورت قطعی در داوری می‌بازد:

$$\text{Length}(y) > \sigma \cdot \ell_0 \implies \text{باخت قطعی در داوری} \tag{۴۲}$$

- **پاداش تقطیر هم‌سیاست چندمعلمه (MOPD):** هدایت مدل دانشجو π_θ توسط ۹ مدل معلم $\pi_{\text{teacher}}^{(d,e)}$ با پاداش متراکم توکنی زیر انجام می‌شود:

$$r_{\text{opd}}^{d}(y_t \mid e, x, y_{<t}) = \text{clip}\left(\text{sg}\left(\log \frac{\pi_{\text{teacher}}^{(d,e)}(y_t \mid x, y_{<t})}{\pi_\theta(y_t \mid e, x, y_{<t})}\right), -R_{\text{max}}, R_{\text{max}}\right) \tag{۴۳}$$

که در آن $\text{sg}(\cdot)$ عملگر توقف گرادیان و R_{max} کران تثبیت است.

- **مدیریت بیات‌شدگی داده‌ها (Data Staleness):** در فرآیند اجرای جزئی (Partial Rollout)، به محض اتمام نسبت $\lambda \in (0,1)$ از کل $N \times K$ مسیر نمونه‌برداری‌شده، داده‌ها وارد بهینه‌سازی سیاست می‌شوند و اثر بیات‌شدگی داده‌های معلق با یک منظم‌سازی موضعی در سطح توکن مهار می‌گردد.

آمار زیرساختی و شواهد تجربی ارزیابی دادگان پس‌آموزش مقیاس تولید داده و کارایی آن در محیط‌های عملیاتی به شرح زیر ثبت شده است:

- در فرآیند پس‌آموزش مدل، در مجموع **۵۱،۲۱۹،۷۴۱** سندباکس در گستره **۱،۵۰۵،۶۷۸** ایمیج سیستمی منحصربه‌فرد اجرا شده است.
- مقیاس‌دهی محاسبات RL افزایش هم‌زمان و معنادار طول گام‌های تعاملی دستیار (Avg. Steps) و نرخ موفقیت (Score %) را در هر ۸ حوزه عملکردی اصلی نشان می‌دهد.
- در آزمون‌های ارزیابی کیفی کورکورانه توسعه وب (Kimi Webdev Bench) در برابر Claude Opus 4.8، مدل **Kimi K3** در مجموع به برتری قاطع **+۰%.۳۱** (با پیروزی در ۶۷.۵۸٪ پرامپت‌ها) و در حوزه تخصصی گرافیک 3D/WebGL به برتری **+۱%.۵۹** دست یافته است (جدول ۳۵).

جدول ۳۵: نتایج داوری کورکورانه کارشناسان در بنچمارک توسعه وب (Kimi Webdev Bench) بر بستر چارچوب Claude Code

حوزه ارزیابی (Domain)	درصد برد (Win)	درصد مساوی (Tie)	درصد باخت (Lose)	تفاضل برد (Win − Lose)
بازی‌ها (Games)	۶%.۵۵	۷%.۳	۷%.۴۰	**+۹%.۱۴**
گرافیک سه‌بعدی و شیدر (3D / WebGL / Shader)	۷%.۷۲	۷%.۱۳	۶%.۱۳	**+۱%.۵۹**
شبیه‌سازی رابط کاربری و سایت (Website / UI Clone)	۶%.۵۲	۱%.۲۱	۳%.۲۶	**+۳%.۲۶**
عملکرد کل (Overall)	**۶%.۵۸**	**۸%.۱۳**	**۶%.۲۷**	**+۰%.۳۱**

مطالعات موردی و توانمندی‌های تعاملی عمیق (Case Studies) شواهد تجربی نشان‌دهنده عمق نفوذ داده‌های آموزشی پس‌آموزش در رفتارهای دنیای واقعی مدل است:

- **بهینه‌سازی بلادرنگ کرنل‌های سخت‌افزاری:** مدل تاخیر اجرای کرنل AttnRes را از ۶.۲۸۳ میلی‌ثانیه به ۴.۱۱۴ میلی‌ثانیه تقلیل داد (**بهبود +۷%.۵۹**) و زمان اجرای KDA را تا ۶%.۷۳ کاهش بخشید.
- **توسعه مستقل کامپایلر MiniTriton:** مدل به صورت خوداتکا یک کامپایلر کامل شبیه Triton به همراه کدهای اسمبلی PTX و زمان اجرای CUDA ساخت که در ضرب ماتریسی به ۹۰% توان تئوری سخت‌افزار (cuBLAS) رسید.
- **طراحی خودکار چیپ Nano-KPU:** در یک اجرای ۴۸ ساعته خودمختار، مدل یک مدار مجتمع کامل سخت‌افزاری را در کتابخانه ۴۵ نانومتری طراحی، شبیه‌سازی و اعتبارسنجی کرد که در فرکانس ۱۰۰ مگاهرتز با توان عملیاتی بیش از ۸،۷۰۰ توکن بر ثانیه کار می‌کند.

جدول ۳۶: ارزیابی جامع عملکرد مدل **Kimi K3** در بنچمارک‌های حوزه‌ای درون‌سازمانی در مقایسه با مدل‌های تجاری و باز

دسته بنچمارک	چارچوب (Harness)	(max) Kimi K3	(max) Fable 5	(max) GPT-5.6 Sol	(max) Opus 4.8	(xhigh) GPT-5.5	(max) GLM-5.2
تجربه و مهارت کدنویسی (Coding Experience)							
Kimi Code Bench 2.0	Claude Code	۷.۷۳	**۹.۷۶**	-	۷.۷۱	-	۲.۶۴
	Kimi Code	۹.۷۲	-	-	-	۰.۶۶	-
	Codex	-	-	۸.۶۴	-	۰.۶۹	-
Coding Experience	Claude Code	**۹.۵۹**	۸.۵۹	-	۰.۵۸	-	۳.۵۳
	Kimi Code	۶.۵۶	-	-	-	-	-
	Codex	-	-	۳.۵۹	-	۸.۵۶	-
تجربه عاملی عمومی (General Agent Experience)							
24/7 ClawBench 2.0	OpenClaw	۳.۴۸	۴.۴۷	**۰.۵۲**	۲.۴۷	۵.۴۸	۲.۴۳
MIRA Bench	MIRA	۱.۶۴	**۹.۷۲**	۲.۶۲	۸.۵۹	۶.۵۴	-
KAET	Kimi Code	۵.۸۳	-	**۴.۸۵**	۷.۷۸	۷.۷۹	۷.۷۴
CLIF Bench	Kimi Code	**۴.۵۲**	-	۶.۵۰	۸.۴۸	۳.۵۲	۲.۳۹
Agentic Vision Bench	Kimi Code	۳.۷۸	۱.۸۱	**۹.۸۲**	۸.۸۲	۹.۷۶	-
Swarm Bench	Kimi Agent	**۳.۷۶**	-	۲.۷۳	۶.۷۲	۸.۶۱	۵.۵۸
Online Experience	Kimi Agent	۹.۷۷	۲.۷۴	**۰.۸۴**	۴.۶۹	۷.۷۳	۰.۶۴
Deep Research Bench	Kimi Agent	**۰.۹۰**	-	۳.۸۵	۲.۸۷	۹.۸۱	۰.۸۴
Finance Bench	N/A	۶.۶۲	-	**۷.۶۲**	۷.۶۰	۴.۵۸	۴.۵۵
KWV Bench	N/A	۷.۶۴	۶.۶۳	**۹.۶۶**	۷.۶۱	۸.۶۵	-
DECK Bench	N/A	۵.۷۳	۰.۷۳	**۷.۷۴**	۹.۶۶	۲.۶۸	۶.۶۸
Agent Behavior Bench	Kimi Work	۰.۶۵	۵.۷۵	**۴.۷۶**	۷.۶۵	۱.۷۰	-
تجربه گفتگو و دقت (Conversational Experience)							
(1 – Hallucination) Faithfulness	N/A	۵.۸۵	-	۸.۸۴	۶.۸۳	**۵.۸۶**	۸.۷۴
Chat All-in-One Bench	Kimi Work	۲.۸۵	**۰.۸۸**	۰.۷۹	۸.۸۳	۸.۷۱	-

- **پژوهش‌های کلان علمی و تحلیلی:** تحلیل ۳۹۱ رویداد امواج گرانشی با سازمان‌دهی هم‌زمان بیش از ۲۰ ساب‌عامل موازی (Swarm)، و همچنین پردازش بیش از ۱۱،۰۰۰ صفحه گزارش مالی جهت تدوین تاریخچه ۴۲ ساله تراشه‌های ASIC.

- **Magistral**

مقدمه و پارادایم داده‌ای پس‌آموزش (Post-Training Data Paradigm) فاز پس‌آموزش (Post-Training) در مدل‌های استدلالی خانواده **Magistral** با هدف ارتقای قابلیت‌های زنجیره تفکر عمیق (Chain-of-Thought - CoT) بدون اتکا به تقطیر از مدل‌های انحصاری پیشین توسعه یافته است [۱۲۵، ۱۰۳]. در این چارچوب، دو مدل با راهبردهای داده‌ای متمایز آموزش دیده‌اند:

- **مدل Magistral Medium:** مبتنی بر وزن‌های پایه Mistral Medium 3 [۱۲۶] که فرایند پس‌آموزش آن **صرفاً با یادگیری تقویتی آنلاین (Pure RLVR)** و بدون هیچ‌گونه داده راه‌اندازی اولیه استدلالی (Cold-Start) به انجام رسیده است.
- **مدل Magistral Small (24B):** مبتنی بر Mistral Small 3 که فرایند پس‌آموزش آن شامل یک فاز تنظیم دقیق تحت نظارت (SFT) با ردپاهای استدلالی پالایش‌شده و سپس یادگیری تقویتی تکمیلی (RL) است.

در این چارچوب، یک دوگانگی بنیادین در ویژگی‌های آماری داده‌ها شناسایی شده است: الگوریتم‌های بهینه‌سازی تقویتی مبتنی بر پاداش‌های قابل راستی‌آزمایی (RLVR) نیازمند مجموعه‌داده‌ای با حجم کم، خلوص فوق‌العاده بالا و سختی بسیار زیاد هستند؛ در حالی که فاز تنظیم نظارتی (SFT) به تنوع گسترده پرامپت‌ها (Prompt Diversity) برای حفظ آنتروپی و توزیع طول استدلال نیاز دارد.

خط‌لوله پالایش و گزینش داده‌های ریاضی (Math Data Curation) داده‌های حوزه ریاضیات طی یک خط‌لوله پالایش دوسطحی شامل فیلترینگ صوری و فیلترینگ سختی مبتنی بر مدل غربال‌گری شدند:

- **پالایش صوری و اعتبارسنجی قطعی (Format Filtering):** از میان مجموعه‌داده اولیه شامل ۶۹۹ هزار نمونه، تمامی مسائل اثباتی کیفی (Proof-based) و چندبخشی که ارزیابی ماشینی آن‌ها با ابزارهای قطعی میسر نبود حذف شدند. همچنین مسائل چندگزینه‌ای به مسائل گزاره‌ای با پاسخ صریح عددی یا عبارتی تبدیل شدند تا احتمال پاسخ‌دهی تصادفی به صفر برسد و امکان

استفاده از موتور محاسبات نمادین SymPy برای انطباق معنایی پاسخ‌ها فراهم گردد. این مرحله حجم داده‌ها را به ۵۰۱ هزار نمونه تقلیل داد.

- **غربال‌گری سختی دو مرحله‌ای در منطقه گلدیلاکس (Goldilocks Filtering):** برای جلوگیری از نمونه‌های بسیار ساده (با مزیت صفر) یا غیرقابل‌حل (بدون سیگنال یادگیری)، فرایند ارزیابی دو مرحله‌ای پیاده‌سازی شد:

۱. در مرحله نخست، با استفاده از مدل Mistral Large 2 [۲۰] به ازای هر مسئله ۱۶ راه‌حل نمونه‌گیری شد و مسائل دارای نرخ موفقیت ۱۰۰٪ یا ۰٪ حذف شدند. داده‌های حاصل برای آموزش یک مدل ۲۴ میلیارد پارامتری ارزیاب سختی به کار رفت.

۲. در مرحله دوم، مدل ارزیاب ۲۴B آموزش‌دیده با RL کل مجموعه ۵۰۱ هزارتایی را مجدداً با ۱۶ پاسخ به ازای هر مسئله باز-ارزیابی کرد تا از حذف اشتباه مسائل دشوار توسط مدل عمومی جلوگیری شود.

- **پالایش بر پایه تضاد اجماع (Consensus-Disagreement Filtering):** مسائلی که در آن‌ها اکثریت پاسخ‌های تولیدی مدل بر سر یک جواب به اجماع می‌رسیدند اما با پاسخ برچسب‌گذاری‌شده (Ground Truth) تضاد داشت، به عنوان خطای برچسب مبنا شناسایی و از مجموعه‌داده حذف شدند:

$$\text{mode}(\{y_i\}_{i=1}^{16}) \neq y_{\text{GT}} \quad \wedge \quad \frac{1}{16}\sum_{i=1}^{16}\mathbb{I}(y_i = \text{mode}) \geq \tau \tag{۴۴}$$

در پایان این فرآیند، تنها **۳۸ هزار مسئله فوق‌العاده خالص و چالش‌برانگیز** برای فاز یادگیری تقویتی گزینش گردید (جدول ۳۷).

جدول ۳۷: مراحل پالایش و حجم دادگان آموزش ریاضی مدل **Magistral**

مرحله فیلترینگ	تعداد داده‌ها	درصد بقا نسبت به کل
داده‌های اولیه خام (Initial Raw Data)	۶۹۹،۰۰۰	۱۰۰.۰%
پس از پالایش صوری و اعتبارسنجی (Format Filtering)	۵۰۱،۰۰۰	۷۱.۶۷%
پس از ارزیابی سختی و اجماع (Difficulty Filtering)	**۳۸،۰۰۰**	**۵.۴۳%**

مهندسی داده‌های کدنویسی و آزمون‌های ارزیابی (Code Data Engineering) پیکره کدنویسی بر پایه مسائل رقابتی برنامه‌نویسی با الزامات اعتبارسنجی زیر طراحی شد:

- **پالایش و اصلاح آزمون‌ها با اجماع اجرا (Test-Case Agreement):** مسائل فاقد آزمون جامع یا راه‌حل‌های مرجع حذف شدند. راه‌حل‌های معتبر روی تمامی آزمون‌ها اجرا شده و آزمون‌های مبهم حذف گردیدند. چنانچه تمامی راه‌حل‌ها روی یک آزمون با خروجی یکسان مردود می‌شدند، خروجی آزمون بر اساس اجماع راه‌حل‌ها اصلاح گردید.
- **تولید آزمون‌های سنتتیک (Synthetic Test Generation):** برای مسائلی که پوشش تست محدودی داشتند، موارد آزمون تکمیلی به‌صورت خودکار تولید و صحت‌سنجی شدند.
- **تکثیر زبانی مسائل (Language Duplication):** مسائل به‌گونه‌ای تنظیم شدند که پیاده‌سازی را به دو زبان پرکاربرد Python و C++20 (با استفاده از هدر از پیش کامپایل‌شده bits/stdc++.h) بطلبند. این فرایند مجموعه‌ای متشکل از **۳۵ هزار مسئله برنامه‌نویسی** تمیز را شکل داد.

چندزبانگی و راهبرد مهار جابه‌جایی زبانی (Multilingual Data Strategy) به منظور جلوگیری از پدیده اختلاط ناخواسته زبان‌ها (Code-Switching) در طول زنجیره تفکر:

- **ترجمه متوازن مسائل:** ۱۰ درصد از کل مسائل ریاضی و کدنویسی انگلیسی به ۶ زبان فرانسوی، اسپانیایی، ایتالیایی، آلمانی، چینی و روسی ترجمه شد.

- **پاداش‌دهی بر پایه طبقه‌بند FastText:** با بهره‌گیری از طبقه‌بند زبانی [۲۱]، سه‌گانه صورت مسئله (Q)، افکار درونی (C) و پاسخ نهایی (A) پس از حذف نشانه‌های ساختاری و کدهای برنامه‌نویسی اعتبارسنجی شد:

$$R_{\text{lang}}(Q, C, A) = \begin{cases} +0.1, & \text{if} \mathcal{C}_{\text{FastText}}(Q) = \mathcal{C}_{\text{FastText}}(C) = \mathcal{C}_{\text{FastText}}(A) \\ 0, & \text{otherwise} \end{cases} \tag{۴۵}$$

داده‌های راه‌اندازی سرد (Cold-Start SFT Data) برای مدل کوچک برای مدل **Magistral Small (24B)**، فاز راه‌اندازی اولیه با ترکیب داده‌های زیر اجرا گردید:

- **ردپاهای استخراج‌شده از Magistral Medium:** راه‌حل‌های صحیح ثبت‌شده در طول آموزش RL مدل بزرگ‌تر، با فیلتر کردن گام‌های ابتدایی و کوتاه، گزینش شدند. برای جلوگیری از سوگیری به سمت مسائل ساده، تعداد تولیدها به ازای هر مسئله محدود شد و مسائل دشوارتر بیش‌نمونه‌گیری (Upsample) شدند.
- **پرامپت‌های متنوع خارجی:** استفاده از مخازن متنی متنوع شامل مجموعه‌های OpenThoughts [۲۲] و OpenR1 [۱۲۷، ۱۲۸] که راه‌حل‌های آن‌ها توسط **Magistral Medium** تولید و غربال شدند.
- **داده‌های عمومی دستورالعمل (General Alignment):** تزریق ۱۰ درصد داده‌های مکالمه عمومی و پیروی از دستورات به پیکره SFT جهت حفظ توانمندی‌های غیر استدلالی نظیر فراخوانی ابزار (Tool Calling).

مدل پایه Mistral Small 3 Instruct برای مدت ۴ دوره (Epochs) روی این ترکیب آموزش دید و سپس وارد فاز RL شد.

ردپاهای متن‌باز در مقیاس بزرگ (Large-Scale OSS Traces) در یک سناریوی پژوهشی مستقل، مدل Mistral Medium 3 با ۳.۱ میلیون ردپای استدلالی برآمده از مدل DeepSeek-R1 (برگرفته از OpenThoughts و OpenR1) تنظیم نظارتی گردید. سپس یادگیری تقویتی صرفاً روی دشوارترین زیرمجموعه مسائل اعمال شد که موجب جهش بیش از ۱۰ واحد درصدی در بنچمارک AIME'25 و ۵ واحد درصدی در LiveCodeBench شد و سطح عملکردی هم‌تراز با مدل‌های پیشرو ایجاد نمود.

تدریج داده‌ای و پویایی طول بافت (Curriculum & Length Dynamics) فرآیند آموزش تقویتی به شیوه برنامه درسی مرحله‌ای (Curriculum) مدیریت شد:

- **پویایی سختی داده‌ها:** همگام با تسلط مدل، مسائل کاملاً حل‌شده از چرخه خارج و مسائل دشوارتر جایگزین گردیدند.
- **برنامه درسی طول تولید:** سقف طول تولید بدون جریمه ($l_{\max} - l_{\text{cache}}$) در سه گام متوالی از 16K به 24K و سپس به 32K توکن افزایش یافت.
- **مقیاس دسته‌ها:** به منظور مدیریت بار حافظه جدول KV-Cache همگام با افزایش طول توالی‌ها، اندازه دسته دنباله‌ها (n_{batch}) در سه مرحله از 8K به 4K و سپس به 2K کاهش داده شد.

فرمولاسیون ریاضی راستی‌آزمایی و پاداش‌دهی داده‌ها (Verifiable Reward Formulation) در خط لوله بهینه‌سازی سیاست گروهی نسبی (GRPO) [۲۴، ۱۲۹]، امتیاز هر نمونه بر اساس رابطه زیر تخصیص می‌یابد:

$$r_i = R_{\text{format}} + R_{\text{correctness}} + R_{\text{lang}} + R_{\text{length}} \tag{۴۶}$$

که در آن پاداش قالب‌بندی $R_{\text{format}} \in \{0, 0.1\}$ شرط ورود به تصحیح پاسخ است. پاداش درستی $R_{\text{correctness}} \in \{0, 0.9\}$ در صورت موفقیت کامل محاسباتی یا عبور از آزمون‌های کدنویسی اعطا می‌شود و تابع پیوسته جریمه طول به فرم زیر محاسبه می‌گردد:

$$R_{\text{length}}(y) = \begin{cases} 0, & |y| \leq l_{\max} - l_{\text{cache}} \\ -0.1 \cdot \frac{|y| - (l_{\max} - l_{\text{cache}})}{l_{\text{cache}}}, & l_{\max} - l_{\text{cache}} < |y| \leq l_{\max} \\ -0.1, & |y| > l_{\max} \end{cases} \tag{۴۷}$$

تعمیم‌پذیری بین‌حوزه‌ای و چندوجهی ناشی از داده‌های متنی (Data Generalization) تحلیل رفتار مدل آموزش‌دیده با داده‌های تفکیک‌شده حقایق تجربی مهمی را آشکار ساخت:

- **انتقال بین ریاضی و کد (Cross-Domain Transfer):** مدل ۲۴B آموزش‌دیده صرفاً با داده‌های ریاضی توانست عملکرد خود در بنچمارک کدنویسی LiveCodeBench را از ۷٪.۲۲ به ۳٪.۳۸ (۶.۱۵+ واحد درصد ارتقای برون‌دامنه) برساند؛ همچنین آموزش صرفاً با داده‌های کد، دقت ریاضی در AIME'24 را ۵.۱۷+ واحد درصد بهبود بخشید (جدول ۳۸).
- **ارتقای رایگان درک چندوجهی (Multimodal Free Lunch):** علی‌رغم این‌که فرایند RL تنها بر روی داده‌های متنی محض اجرا شد، مدل به واسطه گسترش زنجیره تفکر توانست عملکرد در آزمون‌های استدلال چندوجهی نظیر MMMU را از ۰٪.۶۵ به ۰٪.۷۰ و در MMMU-Pro (Vision) را از ۷٪.۳۹ به ۱٪.۵۲ (۴.۱۲+ واحد درصد ارتقا) برساند [۱۳۰، ۱۳۱].

جدول ۳۸: تعمیم‌پذیری بین‌حوزه‌ای در آموزش‌های تک‌دامنه‌ای مدل **Mistral Small 24B**

پیکربندی آموزش داده	AIME'24 (Pass@1)	LiveCodeBench v5 (Pass@1)
چک‌پوینت اولیه (Starting Checkpoint)	۲٪.۳۲	۷٪.۲۲
آموزش صرفاً با داده‌های ریاضی (RL Math only)	**۵٪.۶۲**	**۳٪.۳۸** (۶٪.۱۵+)
آموزش صرفاً با داده‌های کدنویسی (RL Code only)	**۷٪.۴۹** (۵٪.۱۷+)	**۷٪.۴۲**

رویکردهای ناموفق در مهندسی پاداش داده‌ها (Negative Results) دو راهبرد پیشنهادی در فاز پردازش پاداش داده‌ها با شکست مواجه شدند:

- **پاداش متناسب برای آزمون‌های کدنویسی (Proportional Rewards):** اعطای پاداش بر اساس نسبت تست‌های پاس‌شده منجر به افت ۲ درصدی در بنچمارک LiveCodeBench و مهار رشد طول زنجیره تفکر شد؛ چرا که سیگنال‌های کاذب به کدهای واجد خطاهای الگوریتمی اختصاص می‌یافت.
- **جریمه آنتروپی در تلفات (Entropy Bonus):** افزودن ترم آنتروپی به تابع زیان پایداری آموزش را برهم زد (افت شدید آنتروپی در ریاضی و انفجار آنتروپی در کد). در مقابل، تنظیم مستقیم حد بالای برش $\varepsilon_{\text{high}}$ در بازه $[0.26, 0.28]$ پایداری بهینه‌سازی را تضمین کرد.

ارزیابی آماری جامع پیکربندی‌های داده‌ای پس‌آموزش جدول ۳۹ عملکرد نهایی مدل‌ها تحت استراتژی‌های گوناگون داده‌های پس‌آموزش را مقایسه می‌کند. این نتایج اثبات می‌کنند که مدل **Magistral Medium** صرفاً با اتکا بر ۳۸ هزار داده ریاضی و ۳۵ هزار داده کدنویسی به جهش چشمگیر در AIME'24 (از ۸٪.۲۶ به ۶٪.۷۳ در سناریوی Pass@1 و ۰٪.۹۰ در سناریوی Maj@64) دست یافته است.

- sda
- sda
- sda
- sda
- sda

جدول ۳۹: مقایسه عملکردی مدل‌های **Magistral** بر اساس پیکربندی‌های مختلف داده‌های پس‌آموزش

مدل	ترکیب داده SFT	ترکیب داده RL	AIME’24 (Pass@1)	AIME’25 (Pass@1)	MATH-500	LCB v5
Mistral Medium 3	داده عمومی چت	فاقد RL	۸%.۲۶	۲%.۲۱	۰%.۹۱	۱%.۲۹
Magistral Medium	فاقد SFT	۳۸k ریاضی + ۳۵k کد	**۶%.۷۳**	**۹%.۶۴**	**۳%.۹۴**	**۴%.۵۹**
Mistral 24B (SFT only)	ردپاهای Medium + مخازن وب	فاقد RL	۴%.۶۵	۶%.۵۵	۲%.۹۳	۲%.۵۲
Mistral 24B (RL only)	فاقد SFT	۳۸k ریاضی + ۳۵k کد	۸%.۶۵	۹%.۵۱	۴%.۹۵	۴%.۴۶
Magistral Small (24B)	ردپاهای Medium + عمومی (۴ دوره)	۳۸k ریاضی + ۳۵k کد	**۷%.۷۰**	**۸%.۶۲**	**۹%.۹۵**	**۸%.۵۵**

۲.۱ روش (Method)

۱.۲.۱ پیش‌آموزش (Pre-Training)

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

۲.۲.۱ پس‌آموزش (Post-Training)

- 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

۲.۲ بردارکننده (Tokenizer)

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

۳.۲ شبکه عصبی (MLP)

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

- sda
- sda
- sda

۴.۲ تابع فعالساز (Activation Function)

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

۵.۲ نرمالسازی (Normalization)

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

۶.۲ مکانیزم توجه (Attention)

- sda
- sda
- sda
- sda

- sda
- sda
- sda
- sda
- sda

۷.۲ ساختار ترنسفورمر (Transformer)

- 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

۲.۳ تابع هزینه (Loss Function)

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

۴ تنظیم هایپرپارامتر (Hyperparameter Tuning)

- 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 [۱۳۲] 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 [۱۳۳] 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 [۱۳۴] 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 [۱۳۵] 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 [۱۳۶] 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 [۱۳۷] 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 [۱۳۸] 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 [۱۳۹] 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 [۱۴۰] 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 [۱۴۱] 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 [۱۴۲] 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 [۱۴۳] 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 [۱۴۴] 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 [۱۴۵] 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 [۱۴۶] 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 [۱۴۷] 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 [۱۴۸] 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 [۱۴۹] 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 [۱۵۰] 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

۷ یادگیری تقویتی (Reinforcement Learning)

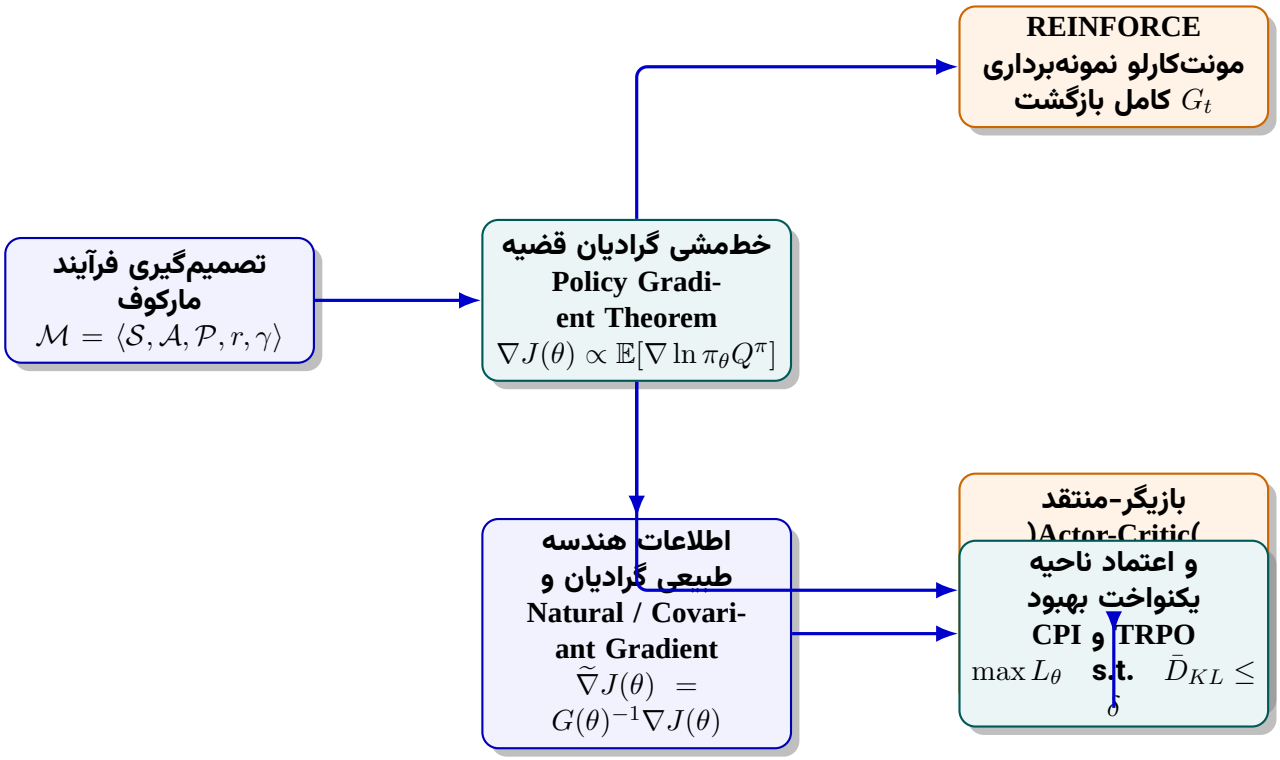
- **مبانی و تئوری الگوریتم‌های مبتنی بر خط‌مشی (Policy-Based RL Theory)**

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 [۱۵۱]:

- **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 [۱۵۲].
- **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 [۱۵۱]).

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:

$$V^\pi(s) \triangleq \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, s_0 = s \right], \tag{۴۸}$$

$$Q^\pi(s,a) \triangleq r(s,a) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s,a) V^\pi(s'). \tag{۴۹}$$

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، $\pi \equiv \pi_\theta$):

$$\begin{aligned}\nabla V^\pi(s) &= \nabla\left[\sum_{a\in\mathcal{A}}\pi(a|s)Q^\pi(s,a)\right]\\ &= \sum_{a\in\mathcal{A}}\left[\nabla\pi(a|s)Q^\pi(s,a)+\pi(a|s)\nabla Q^\pi(s,a)\right].\end{aligned} \tag{۵۰}$$

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

$$\nabla Q^\pi(s,a) = \nabla\left[r(s,a)+\gamma\sum_{s'}\mathcal{P}(s'|s,a)V^\pi(s')\right] = \gamma\sum_{s'}\mathcal{P}(s'|s,a)\nabla V^\pi(s'). \tag{۵۱}$$

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

$$\begin{aligned}\nabla V^\pi(s) &= \sum_a \nabla\pi(a|s)Q^\pi(s,a)+\gamma\sum_{s'}P_\pi(s'|s)\nabla V^\pi(s')\\ &= \sum_{x\in\mathcal{S}}\sum_{t=0}^{\infty}\gamma^t\Pr(s\to x,t,\pi)\sum_a\nabla\pi(a|x)Q^\pi(x,a).\end{aligned} \tag{۵۲}$$

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 [۱۵۳]:

$$\nabla J(\theta) = \sum_{s\in\mathcal{S}}\rho_\pi(s)\sum_{a\in\mathcal{A}}\nabla\pi_\theta(a|s)Q^\pi(s,a) = \mathbb{E}_{s\sim\rho_\pi,a\sim\pi_\theta}\left[\nabla_\theta\ln\pi_\theta(a|s)Q^\pi(s,a)\right]. \tag{۵۳}$$

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 [۱۵۱]:

$$\nabla J(\theta) = \nabla r(\pi) = \sum_{s\in\mathcal{S}}\mu(s)\sum_{a\in\mathcal{A}}\nabla\pi_\theta(a|s)Q^\pi(s,a). \tag{۵۴}$$

الگوریتم REINFORCE و خط مبنا (Baseline)

الگوریتم REINFORCE [۱۵۴] از بازگشت تجربی مسیر $G_t = \sum_{k=t+1}^{T}\gamma^{k-t-1}R_k$ به عنوان برآوردگر نااریب $Q^\pi(S_t,A_t)$ استفاده می‌کند:

$$\theta_{t+1} = \theta_t + \alpha\gamma^t G_t\nabla_\theta\ln\pi_\theta(A_t|S_t). \tag{۵۵}$$

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

$$\sum_{a\in\mathcal{A}}\nabla\pi_\theta(a|s)b(s) = b(s)\nabla_\theta\sum_{a\in\mathcal{A}}\pi_\theta(a|s) = b(s)\nabla_\theta(1) = \mathbf{0}, \tag{۵۶}$$

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 [۱۵۱].

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 [۱۵۵] منیفلد آماری توزیع کل مسیرها (Path Space) را تعریف کردند:

$$P(\xi;\theta) = \rho_0(s_0)\prod_{t=0}^{T-1}\pi_\theta(a_t|s_t)\mathcal{P}(s_{t+1}|s_t,a_t). \tag{۵۷}$$

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 [۱۵۵، ۱۵۶]:

$$G(\theta) = \mathbb{E}_{\xi\sim P(\cdot;\theta)}\left[\nabla_\theta \ln P(\xi;\theta)\nabla_\theta \ln P(\xi;\theta)^\top\right] = \mathbb{E}_{s\sim d_{\rho_0,\pi}}\left[\sum_{a\in\mathcal{A}}\pi_\theta(a|s)\nabla_\theta \ln \pi_\theta(a|s)\nabla_\theta \ln \pi_\theta(a|s)^\top\right]. \tag{۵۸}$$

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

$$\widetilde{\nabla} J(\theta) = G(\theta)^{-1}\nabla J(\theta). \tag{۵۹}$$

کاکاده [۱۵۶] 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 [۱۵۲] اتحاد تفاوت عملکرد (Performance Difference Lemma) را برای هر دو خط‌مشی دلخواه π و $\tilde{\pi}$ اثبات کردند:

$$\eta(\tilde{\pi}) - \eta(\pi) = \sum_{s\in\mathcal{S}}\rho_{\tilde{\pi}}(s)\sum_{a\in\mathcal{A}}\tilde{\pi}(a|s)A^\pi(s,a). \tag{۶۰}$$

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

$$L_\pi(\tilde{\pi}) \triangleq \eta(\pi) + \sum_{s\in\mathcal{S}}\rho_\pi(s)\sum_{a\in\mathcal{A}}\tilde{\pi}(a|s)A^\pi(s,a). \tag{۶۱}$$

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 [۱۵۲]:

$$\eta(\pi_{\text{new}}) \geq L_\pi(\pi_{\text{new}}) - \frac{2\gamma\epsilon}{(1-\gamma)^2}\alpha^2, \quad \text{where} \quad \epsilon = \max_s |\mathbb{E}_{a\sim\pi'}[A^\pi(s,a)]|. \tag{۶۲}$$

بهینه‌سازی در ناحیه اعتماد (TRPO)

شولمن و همکاران [۱۵۷] 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:

$$\eta(\pi_{\text{new}}) \geq L_{\pi_{\text{old}}}(\pi_{\text{new}}) - \frac{4\epsilon\gamma}{(1-\gamma)^2}\max_s D_{TV}(\pi_{\text{old}}(\cdot|s) \parallel \pi_{\text{new}}(\cdot|s))^2. \tag{۶۳}$$

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:

$$\max_{\theta} L_{\theta_{\text{old}}}(\theta) \quad \text{to subject} \quad \bar{D}_{KL}^{\rho_{\theta_{\text{old}}}}(\theta_{\text{old}}, \theta) \leq \delta. \tag{۶۴}$$

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 [۱۵۷]:

$$\theta_{\text{new}} = \theta_{\text{old}} + \sqrt{\frac{2\delta}{g^\top A^{-1} g}} A^{-1} g, \quad g = \nabla_\theta L_{\theta_{\text{old}}}(\theta), \quad A = \nabla_\theta^2 \bar{D}_{KL}. \tag{۶۵}$$

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

$$T_{\pi,\Omega} V \triangleq T_\pi V - \Omega(\pi) = r_\pi - \Omega(\pi) + \gamma P_\pi V, \tag{۶۶}$$

$$T_{*,\Omega} V \triangleq \max_{\pi \in \Delta(\mathcal{A})} T_{\pi,\Omega} V = \Omega^*(q), \quad \text{where} \quad q(s,a) = r(s,a) + \gamma \mathbb{E}_{s'}[V(s')]. \tag{۶۷}$$

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:

$$\pi_{k+1} = \text{argmax}_\pi \left(\langle q_k, \pi \rangle - D_\Omega(\pi \parallel \pi_k) \right). \tag{۶۸}$$

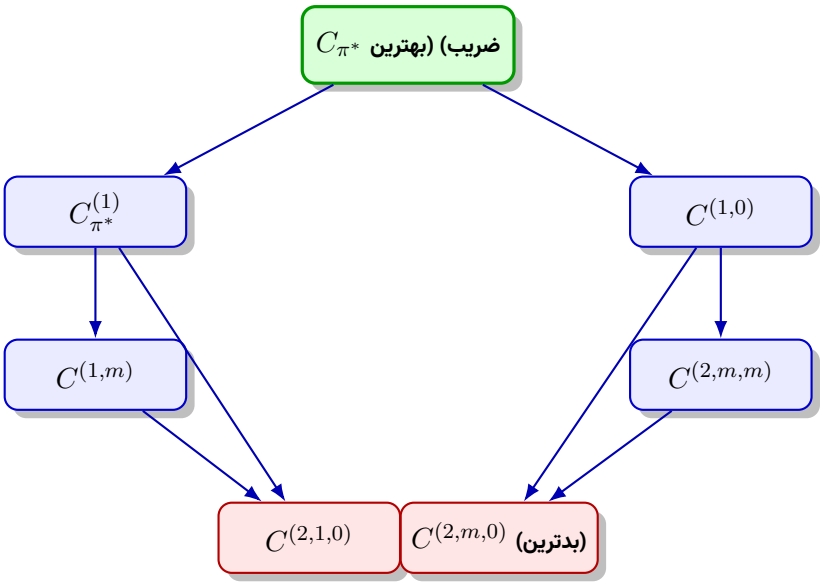
با استفاده از **اتحاد سه‌نقطه‌ای برگمن (Three-Point Identity)**، نرخ همگرایی میانگین حسرت (Regret) به فرم زیر اثبات می‌شود [۱۵۸]:

$$\frac{1}{K} \sum_{k=1}^{K} \|V^* - V^{\pi_k}\|_\infty \leq \frac{1-\gamma^K}{(1-\gamma)^2} \left[\frac{2\gamma \|V^* - V_0\|_\infty + \max_\pi \|D_\Omega(\pi \parallel \pi_0)\|_\infty}{K} \right] = \mathcal{O}\left(\frac{1}{K}\right). \tag{۶۹}$$

مقایسه ساختاریافته الگوریتم‌های تکرار خط‌مشی تقریبی

شرر [۱۵۹] 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 ۴۰: مقایسه مشخصه‌های نظری الگوریتم‌های تکرار خط‌مشی تقریبی [۱۵۹]

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 [۱۵۹، ۱۶۰].

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 [۱۶۱]، 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** [۱۶۲] است. این روش‌ها گرچه در فضاهای عمل گسسته و محیط بازی‌های آتاری [۱۶۳] موفق ظاهر شدند، اما امکان تعمیم کارآمد به فضاهای عمل پیوسته (Continuous Action Spaces) با ابعاد بالا در رباتیک [۱۶۴، ۱۶۵] 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 [۱۵۴] و **A2C/A3C** [۱۶۶] 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 [۱۵۲] و سپس شولمن و همکاران با معرفی الگوریتم **TRPO** [۱۵۷] 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 [۱۶۷] 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:

$$P(\tau;\theta) = \rho_0(s_0) \prod_{t=0}^{T-1} \pi_\theta(a_t|s_t)\mathcal{P}(s_{t+1}|s_t,a_t) \tag{۷۰}$$

هدف بهینه‌سازی، بیشینه‌سازی امید ریاضی بازده کل است:

$$J(\theta) = \mathbb{E}_{\tau\sim P(\cdot;\theta)}[R(\tau)] = \int_\tau P(\tau;\theta)R(\tau)\,d\tau \tag{۷۱}$$

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 ($\nabla_\theta P = P \nabla_\theta \log P$):

$$\begin{aligned}\nabla_\theta J(\theta) &= \nabla_\theta \int P(\tau;\theta) R(\tau)\, d\tau = \int \nabla_\theta P(\tau;\theta) R(\tau)\, d\tau \\ &= \int P(\tau;\theta) \nabla_\theta \log P(\tau;\theta) R(\tau)\, d\tau = \mathbb{E}_{\tau\sim\pi_\theta}\left[\nabla_\theta \log P(\tau;\theta) R(\tau)\right]\end{aligned} \tag{۷۲}$$

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:

$$\hat{g} = \hat{\mathbb{E}}_t \left[\nabla_\theta \log \pi_\theta(a_t|s_t) \hat{A}_t\right] \tag{۷۳}$$

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) [۱۵۲] تعریف می‌شود:

$$L^{CPI}(\theta) = \hat{\mathbb{E}}_t \left[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)} \hat{A}_t\right] = \hat{\mathbb{E}}_t \left[r_t(\theta) \hat{A}_t\right] \tag{۷۴}$$

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:

$$L^{\text{CLIP}}(\theta) = \hat{\mathbb{E}}_t \left[\min\left(r_t(\theta)\hat{A}_t,\ \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\right)\right] \tag{۷۵}$$

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:

$$L^{\text{KLPEN}}(\theta) = \hat{\mathbb{E}}_t \left[r_t(\theta)\hat{A}_t - \beta\, D_{\text{KL}}\left(\pi_{\theta_{\text{old}}}(\cdot|s_t) \parallel \pi_\theta(\cdot|s_t)\right)\right] \tag{۷۶}$$

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:

$$\beta \leftarrow \begin{cases} \beta/2 & \text{if} d < d_{\text{targ}}/1.5 \\ \beta \times 2 & \text{if} d > d_{\text{targ}} \times 1.5 \end{cases} \tag{۷۷}$$

برآوردگر مزیت تعمیم‌یافته (GAE) و تابع هزینه یکپارچه

برای کاهش واریانس، از برآوردگر مزیت تعمیم‌یافته (GAE) منقطع [۱۶۸] استفاده می‌شود:

$$\hat{A}_t = \sum_{l=0}^{T-t-1} (\gamma\lambda)^l \delta_{t+l}^V \quad \text{where} \quad \delta_t^V = r_t + \gamma V(s_{t+1}) - V(s_t) \tag{۷۸}$$

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

$$L_t^{\text{CLIP+VF+S}}(\theta) = \hat{\mathbb{E}}_t \left[L_t^{\text{CLIP}}(\theta) - c_1 \left(V_\theta(s_t) - V_t^{\text{targ}} \right)^2 + c_2 S[\pi_\theta](s_t) \right] \tag{۷۹}$$

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:

$$D_{\text{KL}}(\pi_{\theta_{\text{old}}} \parallel \pi_\theta) \approx \frac{1}{2}(\theta - \theta_{\text{old}})^T F (\theta - \theta_{\text{old}}) \tag{۸۰}$$

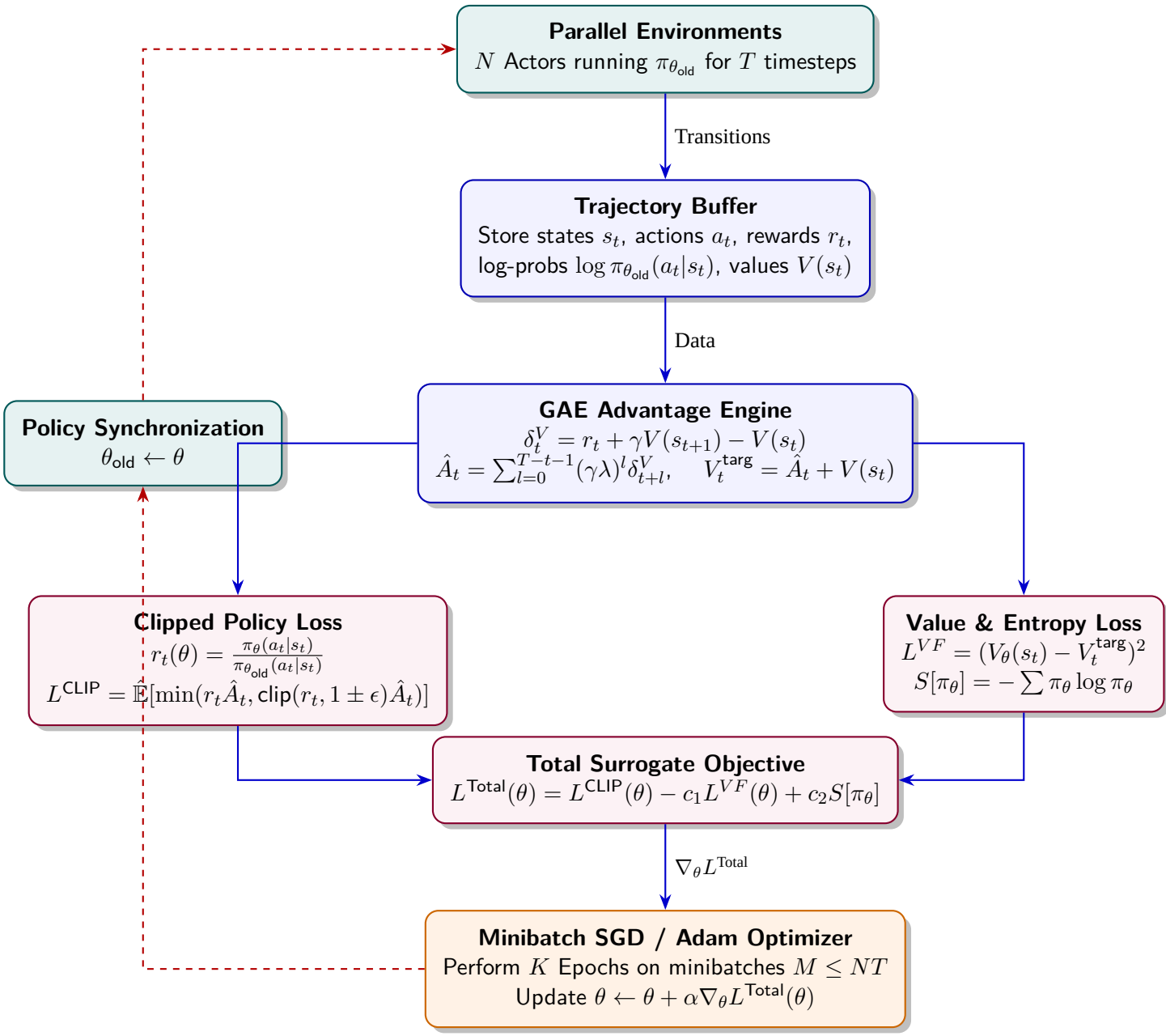
که در آن $F = \mathbb{E}[\nabla_\theta \log \pi_\theta \nabla_\theta \log \pi_\theta^T]$ ماتریس اطلاعات فیشر است. تابع برش در **PPO** تضمین می‌کند که فاصله روی منیفلد ریمانی توزیع‌ها بدون نیاز به تشکیل و معکوس کردن صریح ماتریس F در محدوده‌ای امن محبوس بماند. همچنین با محدود کردن $r_t(\theta) \in [1-\epsilon, 1+\epsilon]$، 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 ۴۱ و روی بازی‌های آتاری در جدول ۴۲ خلاصه شده است:

ابرپارامترهای استاندارد گزارش‌شده در مقاله

جداول ۴۳ و ۴۴ 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 [۱۶۹] تا هم‌ترازسازی مدل‌های زبانی بزرگ از طریق بازخورد انسانی (RLHF) شناخته می‌شود.

جدول ۴۱: مقایسه عملکرد انواع توابع هدف جایگزین در ۷ محیط پیوسته MuJoCo (برگرفته از مقاله مرجع [۱۶۱]).

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 ۴۲: تعداد بازی‌های برده‌شده توسط هر الگوریتم در بنچمارک ۴۹ تایی آتاری (ALE) [۱۶۱].

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 ۴۳: 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 ۴۴: 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 [۲۴]، 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**) [۱۶۱] به عنوان استاندارد اصلی در مرحله یادگیری تقویتی از بازخورد انسانی (RLHF) [۱۷۰] 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:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{\substack{q\sim P(Q) \\ \{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(O|q)}} \left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left(\min\left(\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|q,o_{i,<t})}\hat{A}_{i,t}, \right.\right.\right.$$
$$\left. \text{clip}\left(\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|q,o_{i,<t})}, 1-\varepsilon, 1+\varepsilon \right)\hat{A}_{i,t} \right)$$
$$\left.\left. - \beta\,\mathbb{D}_{\text{KL}}\left(\pi_\theta \parallel \pi_{\text{ref}}\right) \right) \right] \tag{۸۱}$$

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 [۱۷۱] به صورت مستقیم در تابع زیان استفاده می‌شود:

$$\mathbb{D}_{\mathsf{KL}}(\pi_\theta \parallel \pi_{\mathsf{ref}}) = \frac{\pi_{\mathsf{ref}}(o_{i,t}|q, o_{i,<t})}{\pi_\theta(o_{i,t}|q, o_{i,<t})} - \log \frac{\pi_{\mathsf{ref}}(o_{i,t}|q, o_{i,<t})}{\pi_\theta(o_{i,t}|q, o_{i,<t})} - 1 \tag{۸۲}$$

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:

$$f'(r) = 1 - \frac{1}{r}, \qquad f''(r) = \frac{1}{r^2} > 0 \quad (\forall r > 0) \tag{۸۳}$$

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:

$$f(1) = 1 - \log(1) - 1 = 0 \implies f(r) \geq 0 \quad (\forall r > 0) \tag{۸۴}$$

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

$$\begin{aligned}
\mathbb{E}_{x\sim\pi_\theta}\left[\frac{\pi_{\mathsf{ref}}(x)}{\pi_\theta(x)} - \log\frac{\pi_{\mathsf{ref}}(x)}{\pi_\theta(x)} - 1\right] &= \sum_x \pi_\theta(x)\left(\frac{\pi_{\mathsf{ref}}(x)}{\pi_\theta(x)} - \log\frac{\pi_{\mathsf{ref}}(x)}{\pi_\theta(x)} - 1\right) \\
&= \sum_x \pi_{\mathsf{ref}}(x) - \sum_x \pi_\theta(x)\log\frac{\pi_{\mathsf{ref}}(x)}{\pi_\theta(x)} - \sum_x \pi_\theta(x) \\
&= 1 + \sum_x \pi_\theta(x)\log\frac{\pi_\theta(x)}{\pi_{\mathsf{ref}}(x)} - 1 = D_{\mathsf{KL}}(\pi_\theta \parallel \pi_{\mathsf{ref}})
\end{aligned} \tag{۸۵}$$

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:

$$\nabla_\theta \mathcal{L}_{i,t}(\theta) = \nabla_\theta\left[\frac{\pi_\theta(o_{i,t})}{\pi_{\theta_{\mathsf{old}}}(o_{i,t})}\hat{A}_{i,t}\right] - \beta\nabla_\theta\left[\frac{\pi_{\mathsf{ref}}(o_{i,t})}{\pi_\theta(o_{i,t})} - \log\frac{\pi_{\mathsf{ref}}(o_{i,t})}{\pi_\theta(o_{i,t})} - 1\right] \tag{۸۶}$$

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

$$\nabla_\theta\left[\frac{\pi_\theta}{\pi_{\theta_{\mathsf{old}}}}\hat{A}_{i,t}\right] = \hat{A}_{i,t}\nabla_\theta \log \pi_\theta(o_{i,t}) \tag{۸۷}$$

$$\nabla_\theta\left[\frac{\pi_{\mathsf{ref}}}{\pi_\theta} - \log\pi_{\mathsf{ref}} + \log\pi_\theta - 1\right] = -\frac{\pi_{\mathsf{ref}}}{\pi_\theta^2}\nabla_\theta\pi_\theta + \frac{1}{\pi_\theta}\nabla_\theta\pi_\theta = \left(1 - \frac{\pi_{\mathsf{ref}}}{\pi_\theta}\right)\nabla_\theta \log\pi_\theta(o_{i,t}) \tag{۸۸}$$

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

$$\nabla_\theta \mathcal{J}_{\mathsf{GRPO}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left(\hat{A}_{i,t} + \beta\left(\frac{\pi_{\mathsf{ref}}(o_{i,t})}{\pi_\theta(o_{i,t})} - 1\right)\right)\nabla_\theta \log\pi_\theta(o_{i,t}|q, o_{i,<t})\right] \tag{۸۹}$$

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

$$GC_{\mathsf{GRPO}}(q, o, t, \pi_{\theta_{\mathsf{rm}}}) = \hat{A}_{i,t} + \beta\left(\frac{\pi_{\mathsf{ref}}(o_{i,t}|q, o_{i,<t})}{\pi_\theta(o_{i,t}|q, o_{i,<t})} - 1\right) \tag{۹۰}$$

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:

$$\hat{A}_{i,t} = \tilde{r}_i = \frac{r_i - \text{mean}(\mathbf{r})}{\text{std}(\mathbf{r}) + 10^{-8}}, \quad \forall t \in \{1, \ldots, |o_i|\} \tag{۹۱}$$

۲. نظارت بر فرآیند (Process Supervision RL with GRPO): مدل پاداش فرایندی (PRM) [۱۷۲، ۱۷۳] به انتهای هر گام استدلال j پاداش $r_i^{\text{index}(j)}$ اختصاص می‌دهد. با نرمال‌سازی کل پاداش‌های مرحله‌ای درون گروه:

$$\tilde{r}_i^{\text{index}(j)} = \frac{r_i^{\text{index}(j)} - \text{mean}(\mathbf{R})}{\text{std}(\mathbf{R}) + 10^{-8}} \tag{۹۲}$$

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

$$\hat{A}_{i,t} = \sum_{\substack{j=1 \\ \text{index}(j) \geq t}}^{K_i} \tilde{r}_i^{\text{index}(j)} \tag{۹۳}$$

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:

$$\nabla_\theta \mathcal{J}_\mathcal{A}(\theta) = \mathbb{E}_{(q,o)\sim\mathcal{D}} \left[\frac{1}{|o|} \sum_{t=1}^{|o|} GC_\mathcal{A}(q, o, t, \pi_{\text{rf}}) \nabla_\theta \log \pi_\theta(o_t | q, o_{<t}) \right] \tag{۹۴}$$

جدول ۴۵ مقایسه ساختاری این روش‌ها را نشان می‌دهد:

جدول ۴۵: تحلیل ساختار روش‌های هم‌ترازسازی در چارچوب یکپارچه گرادیان [۲۴].

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 [۱۷۴]	$q \sim P_{\text{sft}}(Q),\ o \sim \pi_{\text{sft}}(O\|q)$	قاعده قطعی (Rule)	$\mathbb{I}(o \text{correct is })$	آفلاین تصادفی
DPO [۱۷۵]	$q \sim P_{\text{sft}}(Q),\ o^+, o^- \sim \pi_{\text{sft}}(O\|q)$	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 [۱۶۱]	$q \sim P_{\text{sft}}(Q),\ o \sim \pi_\theta(O\|q)$	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

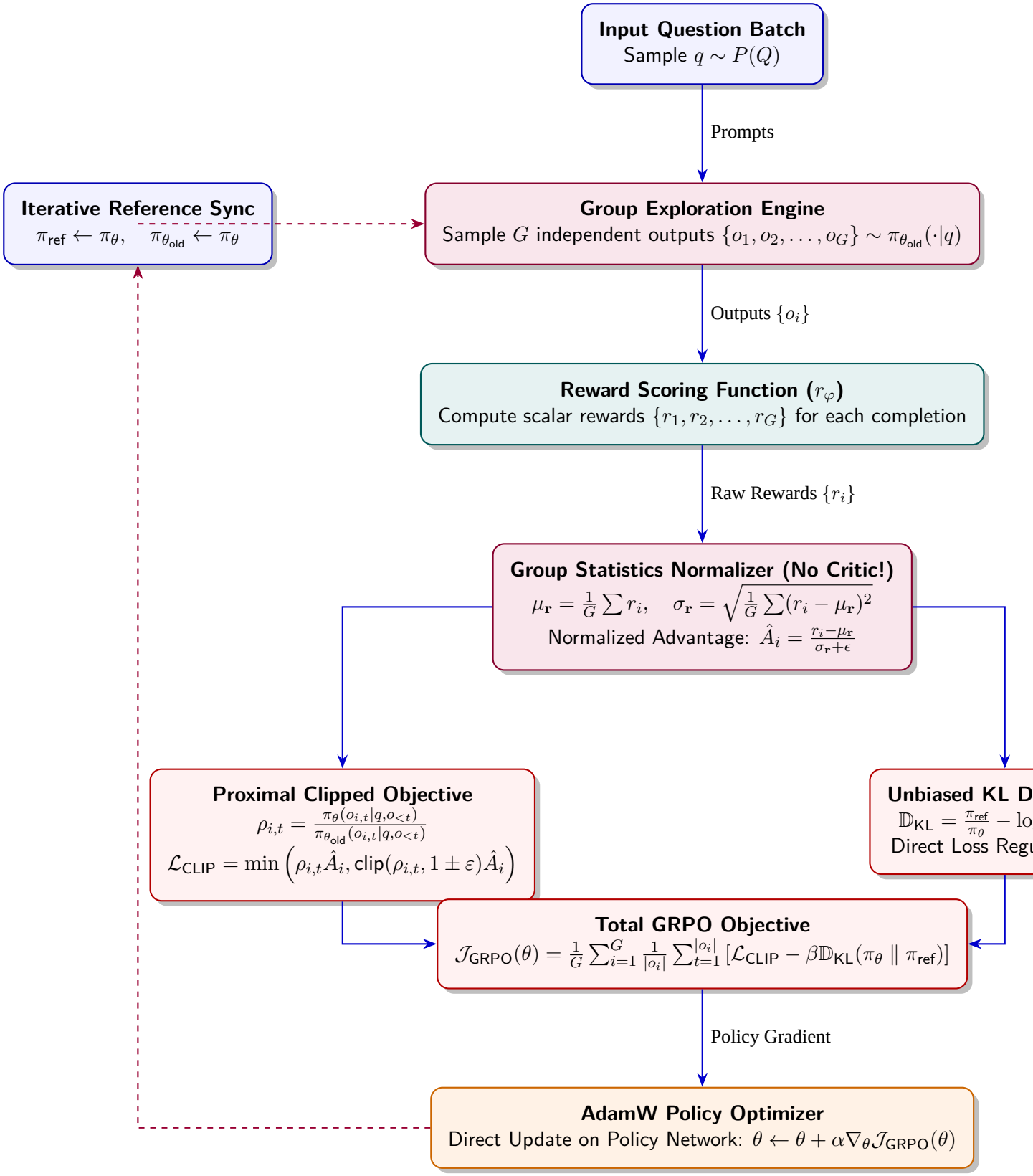
شبه‌کد کامل فرایند بهینه‌سازی سیاست نسبی گروهی تکرارشونده در الگوریتم ۳ تشریح شده است:

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 ۴۶ ارائه شده است:



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

جدول ۴۶: مقایسه عملکرد **DeepSeekMath-RL (GRPO)** با مدل‌های پیشرو در بنچمارک‌های استدلال ریاضی [۲۴].

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 ۴۷ پیکربندی ابرپارامترهای بهینه پیاده‌سازی **GRPO** را نشان می‌دهد:

جدول ۴۷: 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 [۱۷۶]، 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** [۲۴]، 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** [۱۷۷، ۱۷۸] به کار گرفته شده است.

تاریخچه، انگیزه و ضرورت پیدایش الگوریتم

با گسترش مدل‌های زبانی استدلالی نظیر OpenAI o1 [۱۷۹] و **DeepSeek-R1** [۱۰۳]، 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 [۱۶۱]:** 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 [۲۴]:** 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 [۱۷۷، ۱۸۰]. 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:

$$\mathbb{E}_{z\sim\pi_{\text{tar}}}[f(z)] = \mathbb{E}_{z\sim\pi_{\text{beh}}}\left[\frac{\pi_{\text{tar}}(z)}{\pi_{\text{beh}}(z)}f(z)\right] \approx \frac{1}{N}\sum_{k=1}^{N}\frac{\pi_{\text{tar}}(z^{(k)})}{\pi_{\text{beh}}(z^{(k)})}f(z^{(k)}) \qquad (۹۵)$$

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:

$$\pi_\theta(y|x) = \prod_{t=1}^{|y|} \pi_\theta(y_t|x, y_{<t}) \tag{۹۶}$$

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

$$\mathcal{J}_{\text{GSPO}}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\{y_i\}_{i=1}^G\sim\pi_{\theta_{\text{old}}}(\cdot|x)} \left[\frac{1}{G}\sum_{i=1}^{G} \min\left(s_i(\theta)\hat{A}_i,\ \text{clip}(s_i(\theta), 1-\varepsilon, 1+\varepsilon)\hat{A}_i\right)\right] \tag{۹۷}$$

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:

$$\hat{A}_i = \frac{r(x, y_i) - \text{mean}\left(\{r(x, y_i)\}_{i=1}^G\right)}{\text{std}\left(\{r(x, y_i)\}_{i=1}^G\right)} \tag{۹۸}$$

نسبت اهمیت توالی $s_i(\theta)$ بر اساس مفهوم درستی‌نمایی توالی با نرمال‌سازی طولی (Length-Normalized Sequence Likelihood) [۱۸۱] تعریف می‌شود:

$$s_i(\theta) \triangleq \left(\frac{\pi_\theta(y_i|x)}{\pi_{\theta_{\text{old}}}(y_i|x)}\right)^{\frac{1}{|y_i|}} = \exp\left(\frac{1}{|y_i|}\sum_{t=1}^{|y_i|} \log \frac{\pi_\theta(y_{i,t}|x, y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t}|x, y_{i,<t})}\right) \tag{۹۹}$$

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:

$$\begin{aligned}\nabla_\theta \mathcal{J}_{\text{GSPO}}(\theta) &= \nabla_\theta\, \mathbb{E}_{x,\{y_i\}_{i=1}^G}\left[\frac{1}{G}\sum_{i=1}^{G} s_i(\theta)\hat{A}_i\right] \\ &= \mathbb{E}_{x,\{y_i\}_{i=1}^G}\left[\frac{1}{G}\sum_{i=1}^{G} \hat{A}_i\, \nabla_\theta s_i(\theta)\right]\end{aligned} \tag{۱۰۰}$$

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:

$$\nabla_\theta s_i(\theta) = s_i(\theta)\, \nabla_\theta \log s_i(\theta) \tag{۱۰۱}$$

با اعمال لگاریتم بر طرفین معادله (۹۹):

$$\log s_i(\theta) = \frac{1}{|y_i|}\sum_{t=1}^{|y_i|} \left[\log \pi_\theta(y_{i,t}|x, y_{i,<t}) - \log \pi_{\theta_{\text{old}}}(y_{i,t}|x, y_{i,<t})\right] \tag{۱۰۲}$$

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

$$\nabla_\theta \log s_i(\theta) = \frac{1}{|y_i|}\sum_{t=1}^{|y_i|} \nabla_\theta \log \pi_\theta(y_{i,t}|x, y_{i,<t}) \tag{۱۰۳}$$

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

$$\nabla_\theta \mathcal{J}_{\text{GSPO}}(\theta) = \mathbb{E}_{x,\{y_i\}_{i=1}^G}\left[\frac{1}{G}\sum_{i=1}^{G}\left(\frac{\pi_\theta(y_i|x)}{\pi_{\theta_{\text{old}}}(y_i|x)}\right)^{\frac{1}{|y_i|}}\hat{A}_i \cdot \frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\nabla_\theta \log \pi_\theta(y_{i,t}|x, y_{i,<t})\right] \tag{۱۰۴}$$

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

$$\nabla_\theta \mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{x,\{y_i\}_{i=1}^G}\left[\frac{1}{G}\sum_{i=1}^{G}\hat{A}_i \cdot \frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\left(\frac{\pi_\theta(y_{i,t}|x, y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t}|x, y_{i,<t})}\right)\nabla_\theta \log \pi_\theta(y_{i,t}|x, y_{i,<t})\right] \tag{۱۰۵}$$

تحلیل مقایسه‌ای دو گرادیان: در **GRPO** (معادله (۱۰۵))، 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 (۱۰۴))، 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:

$$\mathcal{J}_{\text{GSPO-token}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\min\left(s_{i,t}(\theta)\hat{A}_{i,t},\ \text{clip}(s_{i,t}(\theta), 1-\varepsilon, 1+\varepsilon)\hat{A}_{i,t}\right)\right] \tag{۱۰۶}$$

که در آن نسبت توکن با ساختار زیر تعریف می‌شود:

$$s_{i,t}(\theta) \triangleq \text{sg}[s_i(\theta)] \cdot \frac{\pi_\theta(y_{i,t}|x, y_{i,<t})}{\text{sg}\left[\pi_\theta(y_{i,t}|x, y_{i,<t})\right]} \tag{۱۰۷}$$

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:

$$\nabla_\theta \mathcal{J}_{\text{GSPO-token}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\left(\frac{\pi_\theta(y_i|x)}{\pi_{\theta_{\text{old}}}(y_i|x)}\right)^{\frac{1}{|y_i|}} \cdot \frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\hat{A}_{i,t}\nabla_\theta \log \pi_\theta(y_{i,t}|x, y_{i,<t})\right] \tag{۱۰۸}$$

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

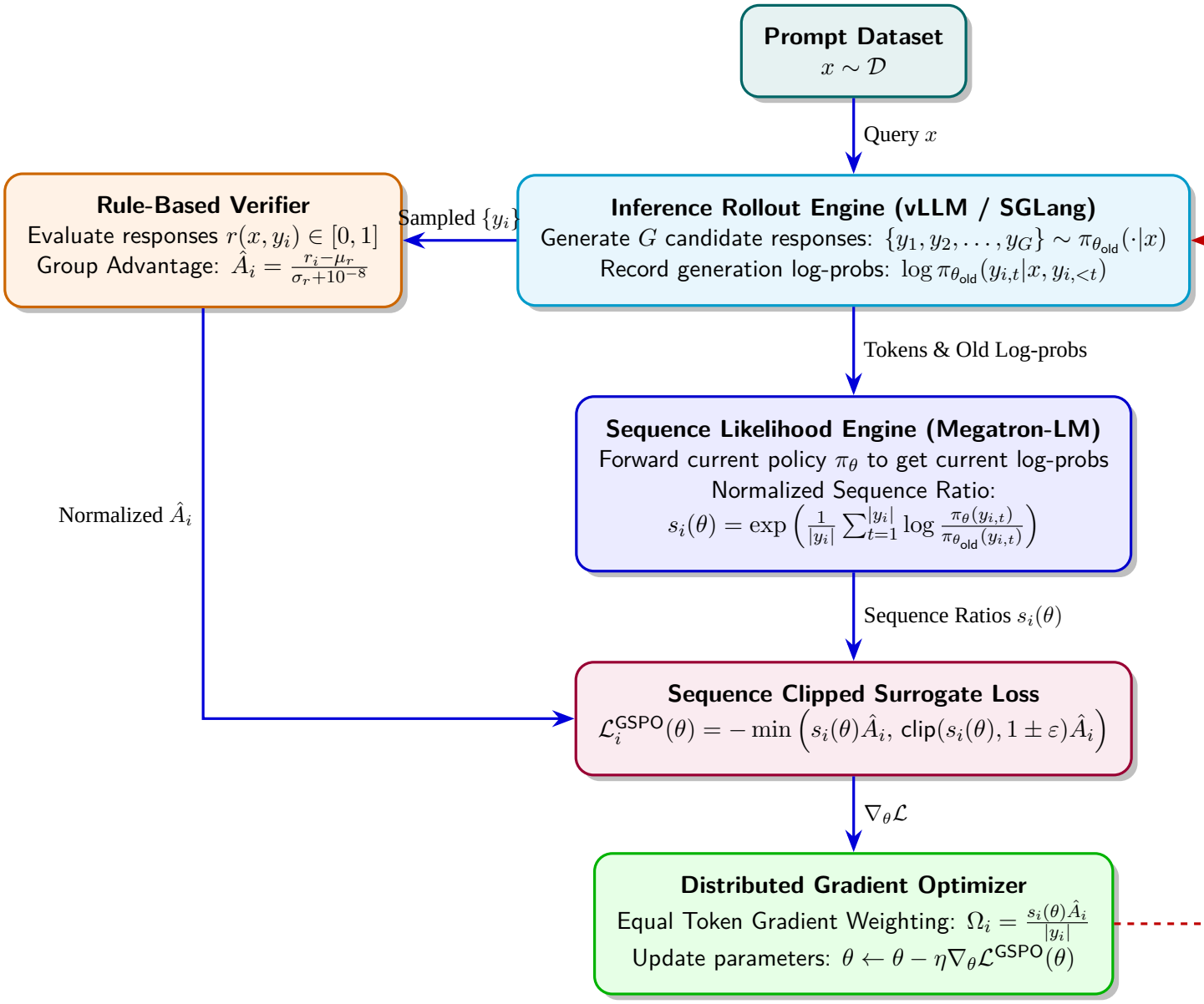
شکل ۳ ساختار جامع خط لوله آموزش، محاسبات نسبت توالی و تفاوت بنیادین جریان داده در الگوریتم **GSPO** در مقایسه با روش‌های پیشین را نمایش می‌دهد.

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 ۴۸ آورده شده است.

جدول ۴۸: مقایسه عملکرد و پایداری الگوریتم‌های **GRPO** و **GSPO** بر روی مدل **Qwen3-30B-A3B** (برگرفته از [۱۷۶]).

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)	۲۷.۲%	۴۳.۴%	۱۹۲۰	نیازمند سربار حافظه
GSPO (رویکرد پیشنهادی)	**۸۲.۶%**	**۶۸.۱%**	**۲۰۶۵**	**پایداری کامل و ذاتی**

پارادوکس کسر توکن‌های برش‌خورده (Clipping Fraction Paradox): جدول ۴۹ مقایسه جالب میانگین توکن‌های برش‌خورده را نشان می‌دهد. با وجود اینکه **GSPO** حدود دو مرتبه بزرگی توکن‌های بیشتری را برش می‌زند (۱۵٪ در برابر ۰.۰۱۳٪)، راندمان نمونه‌ای بسیار بالاتری ارائه می‌دهد؛ امری که نشان می‌دهد گرادیان‌های توکنی **GRPO** سراسر نویز مخرب بوده‌اند.

جدول ۴۹: مقایسه کسر توکن‌های برش‌خورده در طول آموزش [۱۷۶].

شاخص مقایسه‌ای	الگوریتم GRPO	الگوریتم GSPO
کسر توکن‌های برش‌خورده (Clipping Fraction)	۰.۰۱۳ (۰.۰۱۳%)	**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 ۵۰ ابرپارامترهای بهینه‌سازی آموزش **GSPO** را در مقایسه با **GRPO** نشان می‌دهد:

جدول ۵۰: 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 [۱۲۹]، 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** [۵۵] 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** [۱۸۲] 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 [۱۷۹] و **DeepSeek-R1** [۱۸۲]، 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** [۲۴] و **PPO** [۱۶۱] 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:

$$J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}[R(\tau)] = \int P(\tau;\theta)R(\tau)\,d\tau \tag{۱۰۹}$$

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

$$\nabla_\theta J(\theta) = \int \nabla_\theta P(\tau;\theta)R(\tau)\,d\tau = \int P(\tau;\theta)\nabla_\theta \log P(\tau;\theta)R(\tau)\,d\tau = \mathbb{E}_{\tau\sim\pi_\theta}\left[\nabla_\theta \log P(\tau;\theta)R(\tau)\right] \tag{۱۱۰}$$

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:

$$\nabla_\theta J(\theta) = \mathbb{E}_{q\sim\mathcal{D},o\sim\pi_\theta(\cdot|q)}\left[\sum_{t=1}^{|o|}\nabla_\theta \log \pi_\theta(o_t|q,o_{<t})\hat{A}_t\right] \tag{۱۱۱}$$

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:

$$r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})} \tag{۱۱۲}$$

در الگوریتم **GRPO** [۲۴]، 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:

$$\hat{A}_{i,t} = \frac{R_i - \text{mean}(\{R_i\}_{i=1}^G)}{\text{std}(\{R_i\}_{i=1}^G) + \epsilon_{\text{var}}} \tag{۱۱۳}$$

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:

$$J_{\text{DAPO}}(\theta) = \mathbb{E}_{(q,a)\sim\mathcal{D},\{o_i\}_{i=1}^G\sim\pi_{\theta_{\text{old}}}}\left[\frac{1}{\sum_{i=1}^G|o_i|}\sum_{i=1}^G\sum_{t=1}^{|o_i|}\min\left(r_{i,t}(\theta)\hat{A}_{i,t},\ \text{clip}\left(r_{i,t}(\theta),1-\varepsilon_{\text{low}},1+\varepsilon_{\text{high}}\right)\hat{A}_{i,t}\right)\right] \tag{۱۱۴}$$

$$\text{s.t.}\quad 0 < \sum_{i=1}^G \mathbb{I}(\text{is_equivalent}(a,o_i)) < G \tag{۱۱۵}$$

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:

$$R_{\text{length}}(y) = \begin{cases} 0 & |y| \le L_{\text{max}} - L_{\text{cache}} \\ \frac{(L_{\text{max}} - L_{\text{cache}}) - |y|}{L_{\text{cache}}} & L_{\text{max}} - L_{\text{cache}} < |y| \le L_{\text{max}} \\ -1 & |y| > L_{\text{max}} \end{cases} \tag{۱۱۶}$$

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.

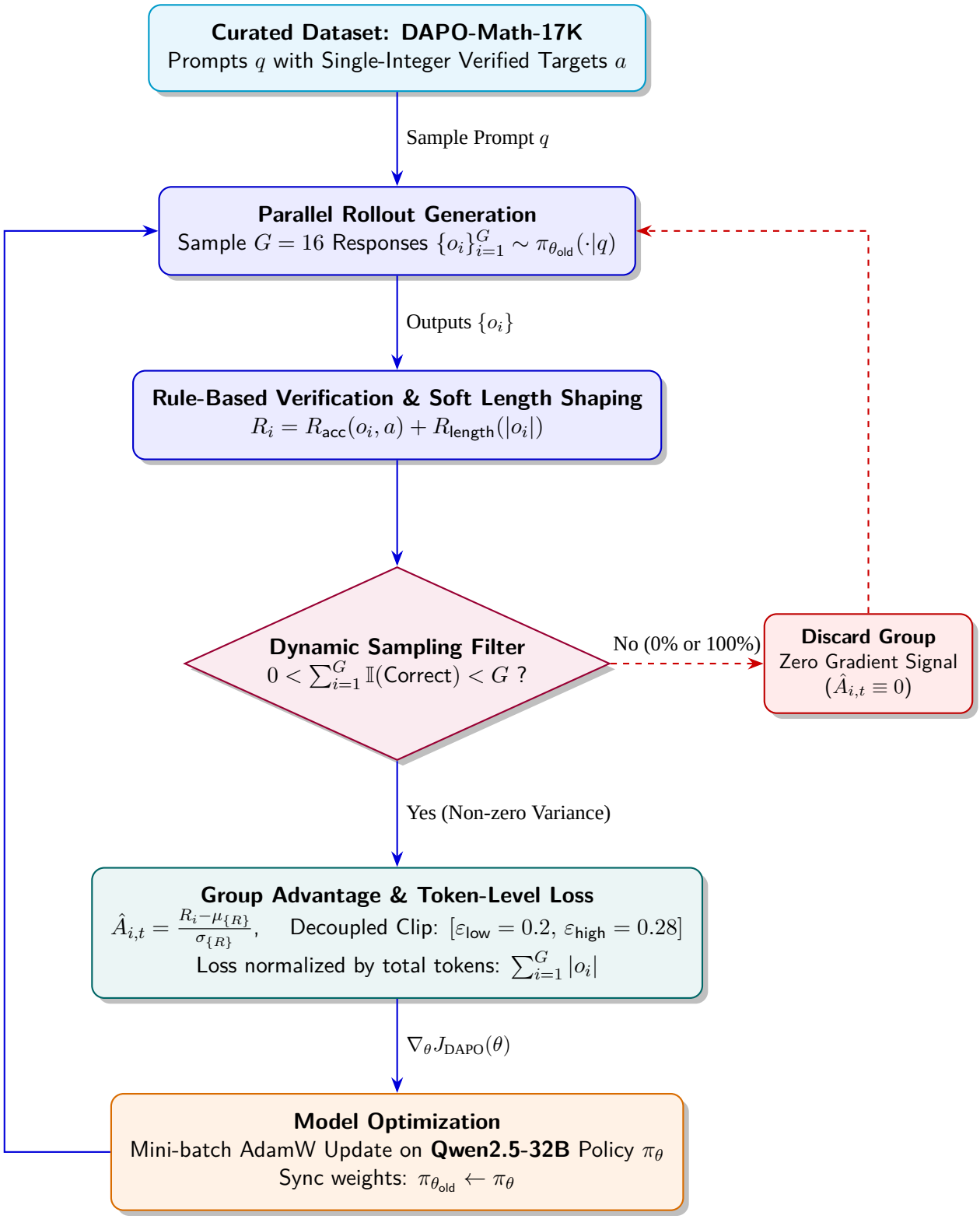
مهندسی داده و تبدیل مجموعه داده (DAPO-Math-17K)

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:

$$R(\hat{y}, y) = \begin{cases} +1 & \text{is_equivalent}(\hat{y}, y) \\ -1 & \text{otherwise} \end{cases} \tag{۱۱۷}$$

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 ۵۱ 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 ۵۱: تحلیل ابلیشن و سهم پیشرونده تکنیک‌های **DAPO** روی مدل **Qwen2.5-32B** در بنچمارک AIME 2024 [۱۲۹].

مدل و راهبرد بهینه‌سازی	دقت AIME 2024 (avg@۳۲)
DeepSeek-R1-Zero-Qwen-32B (مرجع پیشین) [۱۸۲]	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 ۵۲ مقادیر دقیق ابرپارامترهای بهینه‌شده برای آموزش **DAPO** را نمایش می‌دهد:

جدول ۵۲: پیکربندی ابرپارامترهای الگوریتم **DAPO** روی مدل **Qwen2.5-32B**.

ابرپارامتر	مقدار
مدل پایه پیش‌آموزش‌دیده	**Qwen2.5-32B-Base**
فریم‌ورک محاسباتی	verl (HybridFlow) [۱۸۳]
بهینه‌ساز	AdamW [۱۸۴]
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 [۱۸۵]، 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** [۲۴] و **GSPO** [۱۸۶]، 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 [۱۷۹، ۱۸۷، ۱۸۸]، 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** [۲۴] 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** [۱۸۶] 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:

$$\pi_\theta(y \mid q) = \prod_{t=1}^{|y|} \pi_\theta(y_t \mid q, y_{<t}) \tag{۱۱۸}$$

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:

$$\hat{A}_{i,t} = \hat{A}_i = \frac{R_i - \text{mean}\left(\{R_j\}_{j=1}^G\right)}{\text{std}\left(\{R_j\}_{j=1}^G\right) + 10^{-8}} \tag{۱۱۹}$$

نسبت اهمیت در سطح توکن به صورت زیر تعریف می‌شود:

$$r_{i,t}(\theta) \triangleq \frac{\pi_\theta(y_{i,t} \mid q, y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t} \mid q, y_{i,<t})} \tag{۱۲۰}$$

تابع هدف بهینه‌سازی در الگوریتم SAPO:

$$J_{\text{SAPO}}(\theta) = \mathbb{E}_{q\sim\mathcal{D},\{y_i\}_{i=1}^G\sim\pi_{\theta_{\text{old}}}(\cdot|q)}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{t=1}^{|y_i|} f_{i,t}\left(r_{i,t}(\theta)\right)\hat{A}_{i,t}\right] \tag{۱۲۱}$$

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

$$f_{i,t}(x) = \sigma\left(\tau_{i,t}(x-1)\right)\cdot\frac{4}{\tau_{i,t}}, \qquad \tau_{i,t} = \begin{cases}\tau_{\text{pos}} & \text{if}\,\hat{A}_{i,t} > 0\\ \tau_{\text{neg}} & \text{if}\,\hat{A}_{i,t} \le 0\end{cases} \tag{۱۲۲}$$

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 (۱۲۱) و استفاده از قاعده زنجیره‌ای:

$$\nabla_\theta\left[f_{i,t}(r_{i,t}(\theta))\hat{A}_{i,t}\right] = f'_{i,t}(r_{i,t}(\theta))\cdot\nabla_\theta r_{i,t}(\theta)\cdot\hat{A}_{i,t} \tag{۱۲۳}$$

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

$$\nabla_\theta r_{i,t}(\theta) = \frac{\nabla_\theta \pi_\theta(y_{i,t} \mid q, y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t} \mid q, y_{i,<t})} = r_{i,t}(\theta)\nabla_\theta \log \pi_\theta(y_{i,t} \mid q, y_{i,<t}) \tag{۱۲۴}$$

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

$$f'_{i,t}(x) = \frac{4}{\tau_{i,t}}\cdot\frac{d}{dx}\left[\sigma\left(\tau_{i,t}(x-1)\right)\right] = \frac{4}{\tau_{i,t}}\cdot\tau_{i,t}\cdot\sigma'\left(\tau_{i,t}(x-1)\right) = 4\sigma'\left(\tau_{i,t}(x-1)\right) \tag{۱۲۵}$$

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:

$$w_{i,t}(\theta) \triangleq f'_{i,t}(r_{i,t}(\theta)) = 4p_{i,t}(\theta)\big(1-p_{i,t}(\theta)\big) \tag{۱۲۶}$$

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:

$$w_{i,t}(\theta) = 4\cdot\frac{1}{4}\text{sech}^2\left(\frac{\tau_{i,t}}{2}(r_{i,t}(\theta)-1)\right) = \text{sech}^2\left(\frac{\tau_{i,t}}{2}(r_{i,t}(\theta)-1)\right) \tag{۱۲۷}$$

بنابراین، فرم نهایی گرادیان خط‌مشی به دست می‌آید:

$$\nabla_\theta J_{\text{SAPO}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{t=1}^{|y_i|} w_{i,t}(\theta)\, r_{i,t}(\theta)\,\nabla_\theta \log \pi_\theta(y_{i,t} \mid q, y_{i,<t})\,\hat{A}_{i,t}\right] \tag{۱۲۸}$$

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

$$p_{i,t} = \sigma(0) = \frac{1}{2} \implies w_{i,t} = 4\left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = 1 \tag{۱۲۹}$$

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:

$$\frac{\partial \left[\log \pi_\theta(y_{i,t})\hat{A}_{i,t}\right]}{\partial z_v} = \frac{\hat{A}_{i,t}}{\pi_\theta(y_{i,t})} \cdot \frac{\partial \pi_\theta(y_{i,t})}{\partial z_v} \tag{۱۳۰}$$

بر اساس ماتریس ژاکوبی تابع Softmax:

$$\frac{\partial \pi_\theta(y_{i,t})}{\partial z_{y_{i,t}}} = \pi_\theta(y_{i,t})\big(1 - \pi_\theta(y_{i,t})\big) \quad \text{(توکن نمونه‌برداری‌شده):اگر } v = y_{i,t} \tag{۱۳۱}$$

$$\frac{\partial \pi_\theta(y_{i,t})}{\partial z_v} = -\pi_\theta(y_{i,t})\pi_\theta(v) \quad \text{(سایر توکن‌ها):اگر } v \neq y_{i,t} \tag{۱۳۲}$$

با جایگذاری روابط فوق در رابطه (۱۳۰)، گرادیان لاجیت‌ها حاصل می‌شود:

$$\frac{\partial \mathcal{L}_{i,t}}{\partial z_v} = \begin{cases} \big(1 - \pi_\theta(y_{i,t})\big)\hat{A}_{i,t} & \text{if} v = y_{i,t} \quad \text{(Sampled Token)} \\ -\pi_\theta(v)\hat{A}_{i,t} & \text{if} v \neq y_{i,t} \quad \text{(Unsampled Tokens)} \end{cases} \tag{۱۳۳}$$

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 $\mu_i = \log s_i(\theta)$:

$$g_{\tau_i}(z_{i,t}) = g_{\tau_i}(\mu_i) + g'_{\tau_i}(\mu_i)(z_{i,t} - \mu_i) + \frac{1}{2}g''_{\tau_i}(\xi_{i,t})(z_{i,t} - \mu_i)^2 \tag{۱۳۴}$$

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

$$\frac{1}{|y_i|}\sum_{t=1}^{|y_i|} g_{\tau_i}(z_{i,t}) = g_{\tau_i}(\mu_i) + \frac{1}{2|y_i|}\sum_{t=1}^{|y_i|} g''_{\tau_i}(\xi_{i,t})(z_{i,t} - \mu_i)^2 \tag{۱۳۵}$$

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

$$g''_\tau(z) = \alpha^2\big(4\mathsf{sech}^2(\alpha z) - 6\mathsf{sech}^4(\alpha z)\big) \implies \sup_z |g''_\tau(z)| = 2\alpha^2 = \frac{\tau^2}{2} \tag{۱۳۶}$$

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

$$D_i(\theta) \triangleq \left|\frac{1}{|y_i|}\sum_{t=1}^{|y_i|} g_{\tau_i}(z_{i,t}) - g_{\tau_i}\big(\log s_i(\theta)\big)\right| \le \frac{1}{2}\sup_z |g''_{\tau_i}(z)|\mathsf{Var}_i(\theta) = \frac{\tau_i^2}{4}\mathsf{Var}_i(\theta) \tag{۱۳۷}$$

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

$$\nabla_\theta J_{\mathsf{SAPO}}(\theta) \approx \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G} \mathsf{sech}^2\left(\frac{\tau_i}{2}\log s_i(\theta)\right)\nabla_\theta \log s_i(\theta)\hat{A}_i\right] \tag{۱۳۸}$$

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:

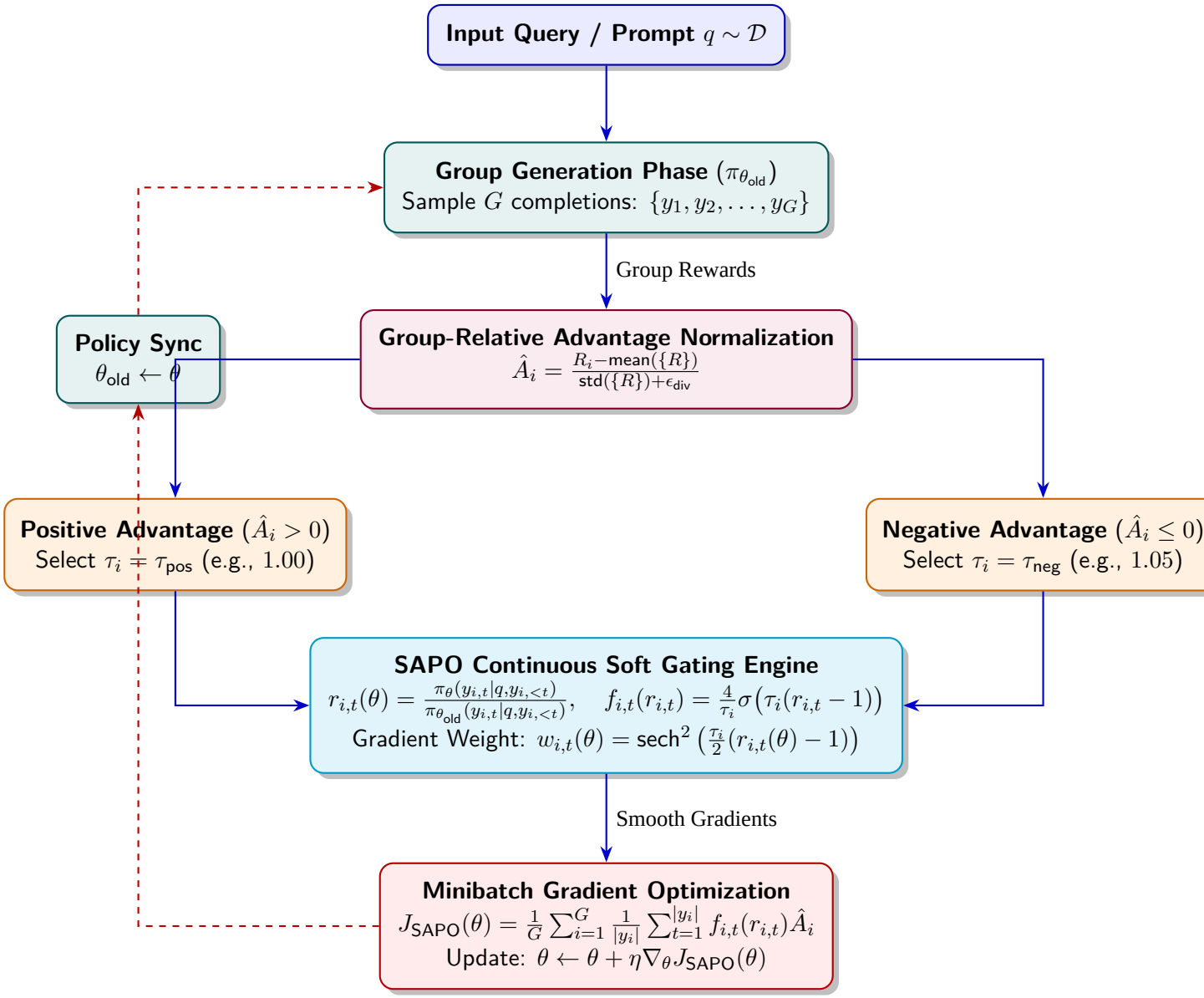
$$D_{\mathsf{KL}}(\pi_{\theta_{\mathsf{old}}} \parallel \pi_\theta) \approx \frac{1}{2}\mathbb{E}\left[(\log r_{i,t}(\theta))^2\right] \tag{۱۳۹}$$

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 ۵۳ خلاصه شده است.

تحلیل ابلیشن دمای نامتقارن (τ_{neg} در برابر τ_{pos}): جدول ۵۴ 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 ۵۳: مقایسه عملکرد نهایی و پایداری در بنچمارک‌های استدلال ریاضی (Pass@1 روی ۱۶ نمونه) [۱۸۵].

الگوریتم	AIME 2025	HMMT 2025	BeyondAIME	وضعیت پایداری آموزش
GRPO-R2 [۲۴، ۱۸۶]	۵۴.۷۶%	۳۰.۵۷%	۲۵.۵۱%	سقوط در گام ۷۵۰ (Collapse)
GSPO [۱۸۶]	۶۰.۷۹%	۴۱.۶۱%	۳۴.۵۴%	سقوط در گام ۱۲۵۰ (Collapse)
SAPO (پیشنهادی)	**۶۲.۸۲%**	**۴۶.۶۴%**	**۳۸.۵۸%**	**کاملاً پایدار (> 1750 گام)**

جدول ۵۴: تحلیل ابلیشن تنظیمات دما در الگوریتم **SAPO** روی مدل Qwen3-30B-A3B [۱۸۵].

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 [۱۸۹] (با ۸ نمونه)، ZebraLogic [۱۹۰] و MathVision [۱۹۱]، 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 [۱۸۰]. 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** [۱۰۳]، 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 [۱۹۲].

با این حال، الگوریتم‌های استاندارد یادگیری تقویتی مانند **PPO** [۱۶۱] و **GRPO** [۲۴] 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 [۱۹۳، ۱۹۴]، ناپایداری بهینه‌سازی و مهار رفتارهای استدلالی بلندمدت می‌شود.

اگرچه الگوریتم **DAPO** [۱۲۹] 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:

$$P(o \mid q; \theta) = \prod_{t=1}^{|o|} \pi_\theta(o_t \mid q, o_{<t}) \tag{۱۴۰}$$

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

$$J(\theta) = \mathbb{E}_{q\sim\mathcal{D}, o\sim\pi_\theta}[R(q,o)] = \sum_q P(q) \int_{\Omega_o} P(o \mid q; \theta) R(q,o)\, do \tag{۱۴۱}$$

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

$$\begin{aligned}
\nabla_\theta J(\theta) &= \sum_q P(q) \int_{\Omega_o} \nabla_\theta P(o \mid q; \theta) R(q,o)\, do \\
&= \sum_q P(q) \int_{\Omega_o} P(o \mid q; \theta) \nabla_\theta \log P(o \mid q; \theta) R(q,o)\, do \\
&= \mathbb{E}_{q\sim\mathcal{D}, o\sim\pi_\theta}\left[\nabla_\theta \log P(o \mid q; \theta) R(q,o)\right]
\end{aligned} \tag{۱۴۲}$$

با اعمال قاعده زنجیره‌ای بر لگاریتم احتمال زنجیره توکن‌ها:

$$\nabla_\theta \log P(o \mid q; \theta) = \sum_{t=1}^{|o|} \nabla_\theta \log \pi_\theta(o_t \mid q, o_{<t}) \tag{۱۴۳}$$

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:

$$\mathbb{E}_{x\sim p}[f(x)] = \int p(x)f(x)\,dx = \int q(x)\frac{p(x)}{q(x)}f(x)\,dx = \mathbb{E}_{x\sim q}\left[\frac{p(x)}{q(x)}f(x)\right] \tag{۱۴۴}$$

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

$$r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})} \tag{۱۴۵}$$

در فرمولاسیون **GRPO** [۲۴]، 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:

$$\hat{A}_{i,t} = \hat{A}_i = \frac{R_i - \text{mean}(\{R_j\}_{j=1}^G)}{\text{std}(\{R_j\}_{j=1}^G) + \delta} \tag{۱۴۶}$$

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:

$$\mathcal{J}_{\text{REINFORCE}}(\theta) = \mathbb{E}_{(q,a)\sim\mathcal{D}, o_i\sim\pi_{\theta_{\text{old}}}}\left[\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\text{sg}(r_{i,t}(\theta))\hat{A}_{i,t}\log\pi_\theta(o_{i,t} \mid q, o_{i,<t})\right] \tag{۱۴۷}$$

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:

$$\hat{r}_{i,t}(\theta) = \text{clip}\left(r_{i,t}(\theta),\ 1-\epsilon_{\text{low}}^{IS},\ 1+\epsilon_{\text{high}}^{IS}\right) \tag{۱۴۸}$$

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

$$\mathcal{J}_{\text{CISPO}}(\theta) = \mathbb{E}_{q\sim\mathcal{D},\{o_i\}_{i=1}^G\sim\pi_{\theta_{\text{old}}}}\left[\frac{1}{\sum_{i=1}^G|o_i|}\sum_{i=1}^G\sum_{t=1}^{|o_i|}\text{sg}\left(\hat{r}_{i,t}(\theta)\right)\hat{A}_{i,t}\log\pi_\theta(o_{i,t} \mid q, o_{i,<t})\right] \tag{۱۴۹}$$

با محاسبه صریح گرادیان نسبت به پارامترهای θ:

$$\nabla_\theta\mathcal{J}_{\text{CISPO}}(\theta) = \mathbb{E}\left[\frac{1}{\sum_{i=1}^G|o_i|}\sum_{i=1}^G\sum_{t=1}^{|o_i|}\hat{r}_{i,t}(\theta)\hat{A}_{i,t}\nabla_\theta\log\pi_\theta(o_{i,t} \mid q, o_{i,<t})\right] \tag{۱۵۰}$$

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:

$$\mathcal{J}_{\text{unify}}(\theta) = \mathbb{E}\left[\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\text{sg}\left(\hat{r}_{i,t}(\theta)\right)\hat{A}_{i,t}\log\pi_\theta(o_{i,t}\mid q, o_{i,<t})M_{i,t}\right] \tag{۱۵۱}$$

که در آن ماسک ناحیه اطمینان برابر است با:

$$M_{i,t} = \begin{cases} 0 & \text{if}\,\hat{A}_{i,t} > 0 \text{ and } r_{i,t}(\theta) > 1+\epsilon_{\text{high}} \\ 0 & \text{if}\,\hat{A}_{i,t} < 0 \text{ and } r_{i,t}(\theta) < 1-\epsilon_{\text{low}} \\ 1 & \text{otherwise} \end{cases} \tag{۱۵۲}$$

جدول ۵۵ تفاوت ساختاری این الگوریتم‌ها را به تصویر می‌کشد.

جدول ۵۵: 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 [۱۶۱]	$r_{i,t}(\theta)$	شرطی (حذف گرادیان در خروج از مرز)	خیر	بله
GRPO [۲۴]	$r_{i,t}(\theta)$	شرطی (حذف گرادیان در خروج از مرز)	خیر	خیر
DAPO [۱۲۹]	$r_{i,t}(\theta)$	شرطی با ϵ_{high} بسیار بزرگتر	تا حدی	خیر
CISPO [۱۸۰]	$\text{clip}(r_{i,t}, 1-\epsilon_{\text{low}}^{IS}, 1+\epsilon_{\text{high}}^{IS})$	**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:

$$\text{Var}\left[\nabla_\theta \mathcal{J}_{\text{CISPO}}\right] \leq (1+\epsilon_{\text{high}}^{IS})^2 \cdot \text{Var}\left[\hat{A}_{i,t}\nabla_\theta \log \pi_\theta\right] \tag{۱۵۳}$$

  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 [۱۹۴].

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) [۱۹۵] 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.

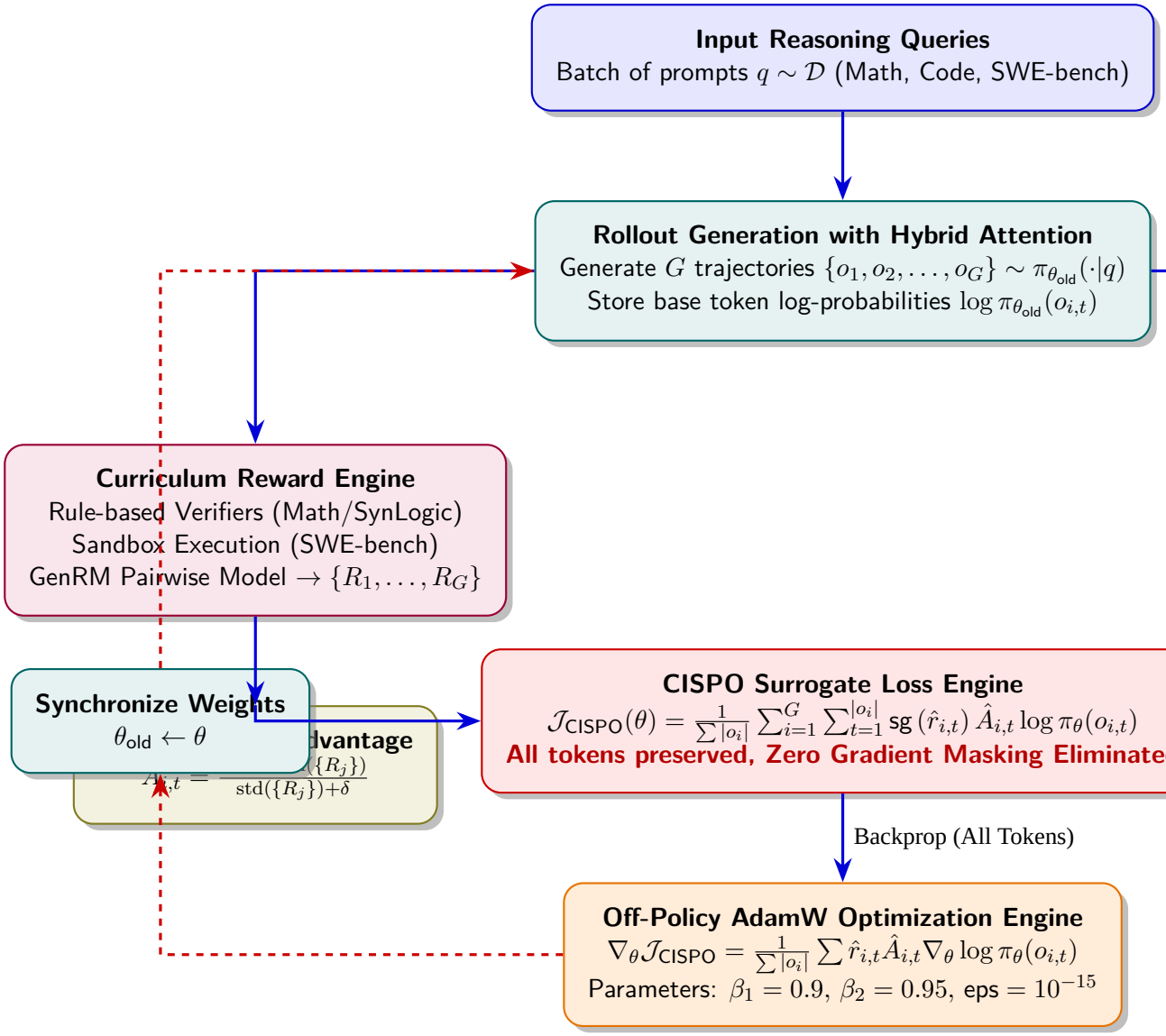
۲. **پیکربندی بهینه‌ساز AdamW:** با توجه به بازه گسترده بزرگی گرادیان‌ها (10^{-18} تا 10^{-5})، پارامترهای بهینه‌ساز به صورت $\beta_1 = 0.9$، $\beta_2 = 0.95$ و $\epsilon = 10^{-15}$ تنظیم شدند [۱۹۶].

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** [۱۸۰]، 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 ۵۶ نتایج جامع مدل‌های آموزش‌دیده با **CISPO** را در مقایسه با پیشرفته‌ترین مدل‌های متن‌باز و تجاری جهان نشان می‌دهد.

جدول ۵۶: ارزیابی مقایسه‌ای مدل‌های **MiniMax-M1** آموزش‌دیده با الگوریتم **CISPO** در بنچمارک‌های استاندارد بین‌المللی [۱۸۰].

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 [۱۹۷]. 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 [۱۹۸، ۱۹۹، ۱۷۰]، 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 [۲۰۰] و سینگهال و همکاران [۲۰۱] 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 [۷۲، ۲۰۲]، اجرای کدهای برنامه‌نویسی و آزمون‌های ساختاریافته دستورپذیری [۲۰۳]، 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** [۲۰۴]، **Quiet-STaR** [۲۰۵]، **TRICE** [۲۰۶]، بازخورد محیطی در کدنویسی [۲۰۷] و **VinePPO** [۲۰۸] 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:

$$\max_{\pi_\theta} \mathcal{J}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi_\theta(\cdot|x)} \left[R_{\text{RLVR}}(x, y)\right] \tag{۱۵۴}$$

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

$$R_{\text{RLVR}}(x, y) = v(x, y) - \beta\, D_{\text{KL}}\left(\pi_\theta(y|x) \parallel \pi_{\text{ref}}(y|x)\right) \tag{۱۵۵}$$

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

$$v(x, y) = \begin{cases} \alpha & \text{اگر استخراج پاسخ با پاسخ صحیح برابر باشد } (\text{Verify}(x, y) = 1) \\ 0 & \text{در غیر این صورت} \end{cases} \tag{۱۵۶}$$

در پیاده‌سازی رسمی [۱۹۷]، 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:

$$D_{\text{KL}}(\pi_\theta(y|x) \parallel \pi_{\text{ref}}(y|x)) = \sum_{t=1}^{T} \log\left(\frac{\pi_\theta(y_t \mid x, y_{<t})}{\pi_{\text{ref}}(y_t \mid x, y_{<t})}\right) \tag{۱۵۷}$$

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

$$r_t(x, y) = \begin{cases} -\beta \log\left(\frac{\pi_\theta(y_t|x, y_{<t})}{\pi_{\text{ref}}(y_t|x, y_{<t})}\right) & \text{for} 1 \le t < T \\ v(x, y) - \beta \log\left(\frac{\pi_\theta(y_T|x, y_{<T})}{\pi_{\text{ref}}(y_T|x, y_{<T})}\right) & \text{for} t = T \end{cases} \tag{۱۵۸}$$

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:

$$\delta_t^V = r_t + \gamma V_\phi(s_{t+1}) - V_\phi(s_t) \tag{۱۵۹}$$

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

$$\hat{A}_t^{\mathsf{GAE}} = \sum_{l=0}^{T-t-1} (\gamma\lambda)^l \delta_{t+l}^V \tag{۱۶۰}$$

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

$$\hat{A}_t^{\mathsf{norm}} = \frac{\hat{A}_t^{\mathsf{GAE}} - \mu(\hat{A}^{\mathsf{GAE}})}{\sigma(\hat{A}^{\mathsf{GAE}}) + 10^{-8}} \tag{۱۶۱}$$

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

$$L^{\mathsf{CLIP}}(\theta) = \hat{\mathbb{E}}_t \left[\min \left(\rho_t(\theta)\hat{A}_t^{\mathsf{norm}},\ \mathsf{clip}(\rho_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t^{\mathsf{norm}} \right) \right] \tag{۱۶۲}$$

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:

$$\mathcal{L}_{\mathsf{total}}(\theta, \phi) = -L^{\mathsf{CLIP}}(\theta) + c_1 L^{\mathsf{VF}}(\phi) \tag{۱۶۳}$$

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 [۱۹۷] 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).

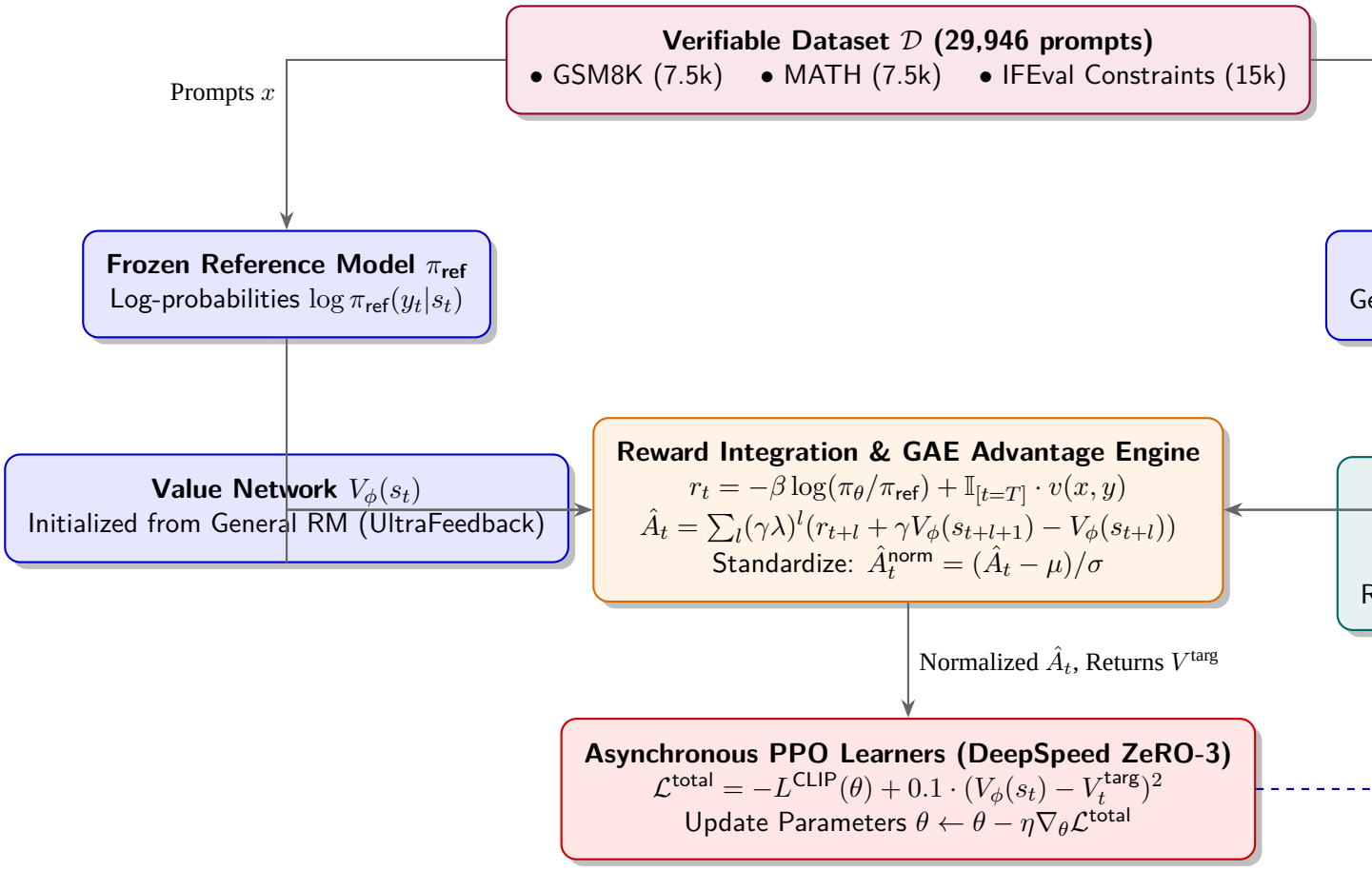
۲. معماری ناهمگام (Asynchronous RL Infrastructure):
برای غلبه بر گلوگاه محاسباتی فاز تولید، موتور استنتاج vLLM [۲۰۹] با مکانیزم PagedAttention بر روی گره‌های اختصاصی تفکیک شده و مدل‌های یادگیرنده (Learners) با استفاده از فناوری DeepSpeed ZeRO-3 [۲۱۰] گرادیان‌ها را به طور همزمان محاسبه می‌کنند [۲۱۱]. 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 [۲۱۲].

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 ۵۷ خلاصه شده است.

جدول ۵۷: ارزیابی مقایسه‌ای مدل‌های آموزش‌دیده با **RLVR** در مقایسه با مبدا **DPO** و خط مبنای **Llama 3.1 Instruct** [۱۹۷].

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

جدول ۵۸: 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 [۲۱۳]، 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) [۲۴] 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 Think-ing 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** [۱۶۱] 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** [۲۴] 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» [۲۱۴] و «اتلاف محاسبات در دسته‌های با گرادیان صفر» مواجه می‌شود [۱۲۹].
- **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 [۲۱۵].

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:

$$\pi_\theta(y|x) = \prod_{t=1}^{T} \pi_\theta(y_t|x, y_{<t}) \tag{۱۶۴}$$

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:

$$J(\theta) = \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(\cdot|x)}[r(x, y)] = \sum_x P(x) \sum_y \pi_\theta(y|x) r(x, y) \tag{۱۶۵}$$

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

$$\begin{aligned}
\nabla_\theta J(\theta) &= \sum_x P(x) \sum_y \nabla_\theta \pi_\theta(y|x) r(x, y) \\
&= \sum_x P(x) \sum_y \pi_\theta(y|x) \nabla_\theta \log \pi_\theta(y|x) r(x, y) \\
&= \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(\cdot|x)} \left[\nabla_\theta \log \pi_\theta(y|x) r(x, y)\right]
\end{aligned} \tag{۱۶۶}$$

با باز کردن رابطه برای تک‌تک توکن‌های توالی پاسخ:

$$\nabla_\theta \log \pi_\theta(y|x) = \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(y_t|x, y_{<t}) \tag{۱۶۷}$$

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

$$\nabla_\theta J(\theta) = \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(\cdot|x)} \left[\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(y_t|x, y_{<t}) \left(r(x,y) - b(x)\right) \right] \tag{۱۶۸}$$

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 [۲۱۴]، 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:

$$A_{i,t} = r(x, y_i) - \text{mean}\left(\{r(x, y_k)\}_{k=1}^{G}\right) = r(x, y_i) - \frac{1}{G}\sum_{k=1}^{G} r(x, y_k) \tag{۱۶۹}$$

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: $\frac{1}{\sum_{i=1}^{G}|y_i|}$ [۱۲۹].

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

$$r_{i,t}(\theta) = \frac{\pi_\theta(y_{i,t}|x, y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t}|x, y_{i,<t})} \tag{۱۷۰}$$

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 [۱۲۹]:

$$\text{clip}\left(r_{i,t}(\theta), 1 - \varepsilon_{\text{low}}, 1 + \varepsilon_{\text{high}}\right) \tag{۱۷۱}$$

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 [۲۱۵]:

$$w_{i,t}^{\text{TIS}} = \min\left(\frac{\pi(y_{i,t}|x, y_{i,<t}; \theta_{\text{old}})}{\pi_{\text{vllm}}(y_{i,t}|x, y_{i,<t}; \theta_{\text{old}})},\ \rho\right) \tag{۱۷۲}$$

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 [۱۲۹، ۲۱۴].

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:

(۱۷۳)

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

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 [۲۱۶] 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 [۲۱۷] 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 [۲۱۸]، 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.

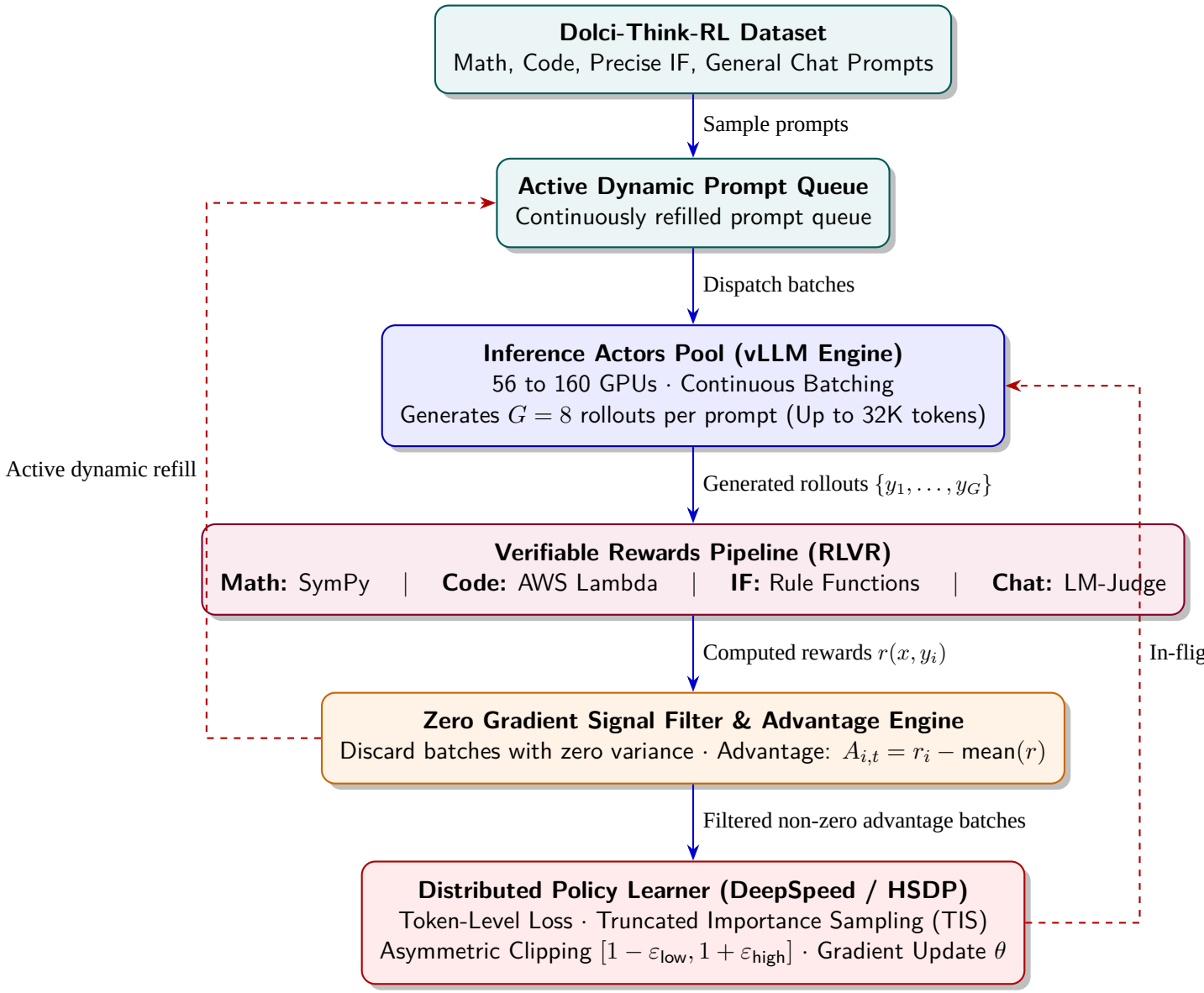
معماری زیرساخت محاسباتی و مهندسی سیستم

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 [۲۱۹]، 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 ۵۹ تاثیر مولفه‌های مهندسی **OlmoRL** بر شتاب محاسباتی و جدول ۶۰ عملکرد مدل پرچم‌دار **Olmo 3 Think-32B** را در مقایسه با برترین مدل‌های هم‌رده نمایش می‌دهند.

ابرپارامترهای استاندارد گزارش‌شده در مقاله

جدول ۶۱ مقادیر دقیق ابرپارامترهای آموزشی الگوریتم **OlmoRL** را برای مدل‌های مختلف نشان می‌دهد:

جدول ۵۹: تحلیل ابلیشن زیرساخت **OlmoRL** و تاثیر آن بر سرعت آموزش و بهره‌وری سخت‌افزار [۲۱۳].

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 ۶۰: مقایسه عملکرد مدل استدلالی **Olmo 3 Think-32B** در بنچمارک‌های استاندارد [۲۱۳].

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 ۶۱: ابرپارامترهای نهایی آموزش یادگیری تقویتی **OlmoRL** در مدل‌های خانواده **Olmo 3** [۲۱۳].

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 ($\varepsilon_{\text{low}}, \varepsilon_{\text{high}}$)	(۲۷۲.۰ ،۲.۰)	(۲۷۲.۰ ،۲.۰)	(۲۷۲.۰ ،۲.۰)	(۲۷۲.۰ ،۲.۰)
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	۵۶ / ۱۶	۱۶۰ / ۶۴	۵۶ / ۸	۶۴ / ۸

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] DeepSeek-AI. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

[2] X. Zhu, D. Cheng, H. Li, K. Zhang, E. Hua, X. Lv, N. Ding, Z. Lin, Z. Zheng, and B. Zhou. How to synthesize text data without model collapse? *arXiv preprint arXiv:2412.14689*, 2024.

[3] H. Ding, Z. Wang, G. Paolini, V. Kumar, A. Deoras, D. Roth, and S. Soatto. Fewer truncations improve language modeling. *arXiv preprint arXiv:2404.10830*, 2024.

[4] Keller Jordan et al. Muon: An optimizer for hidden layers in neural networks. https://kellerjordan.github.io/posts/muon/, 2024.

[5] J. Liu, J. Su, X. Yao, Z. Jiang, G. Lai, Y. Du, et al. Muon is scalable for llm training. *arXiv preprint arXiv:2502.16982*, 2025.

[6] Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

[7] GLM-5 Team, Zhipu AI, and Tsinghua University. Glm-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026.

[8] Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, et al. Datacomp-lm: In search of the next generation of training sets for language models. *arXiv preprint arXiv:2406.11794*, 2025.

[9] Chang Gao, Xidong Wu, Zihan Lin, Dong Zhang, and Shengding Hu. Nextlong: Toward effective long-context training without long documents. *arXiv preprint arXiv:2502.00933*, 2025.

[10] Jun Jia, Zihan Chen, Xidong Wu, Chang Gao, Zihan Lin, Dong Zhang, Shengding Hu, and Baining Guo. Entropylong: Effective long-context training via predictive uncertainty. *arXiv preprint arXiv:2502.15840*, 2025.

[11] OpenAI. Multi-round context recall (mrcr) dataset. https://huggingface.co/datasets/openai/mrcr, 2024.

[12] DeepSeek-AI, Aixin Liu, Alex Mei, et al. Deepseek-v3.2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*, 2025.

[13] Kimi Team. Kimi k3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026.

[14] Kimi Team. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.

[15] Kimi Team. Kimi k2.5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026.

[16] Kimi Team. Kimi linear: An expressive, efficient attention architecture. *arXiv preprint arXiv:2510.26692*, 2025.

[17] Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models. *arXiv preprint arXiv:2309.00071*, 2023.

[18] Shengding Hu, Yuge Tu, Xu Han, et al. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024.

[19] Mistral AI. Magistral: Mistral's first reasoning model. *arXiv preprint arXiv:2506.10910*, 2025.

[20] Mistral AI. Mistral large 2. https://mistral.ai/news/mistral-large-2407, 2024.

[21] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*, 2016.

[22] Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, et al. Openthoughts: Data recipes for reasoning models. *arXiv preprint arXiv:2506.04178*, 2025.

[23] Hugging Face. Open r1: A fully open reproduction of deepseek-r1. https://github.com/huggingface/open-r1, 2025.

[24] 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.

[25] MiniMax. The minimax-m2 series: Mini activations unleashing max real-world intelligence. *arXiv preprint arXiv:2605.26494*, 2026.

[26] Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding. In *International Conference on Machine Learning (ICML)*, pages 19274–19286, 2023.

[27] Fabian Gloeckle, Badr Youbi Idrissi, Baptiste Rozière, David Lopez-Paz, and Gabriel Synnaeve. Better & faster large language models via multi-token prediction. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

[28] Damai Dai, Chengqi Deng, Chenggang Zhao, R. X. Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y. Wu, et al. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 2698–2717, 2024.

[29] 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. In *NeurIPS Datasets and Benchmarks Track*, 2021.

[30] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*, 2021.

[31] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

[32] Kaijing Ma, Xinrun Du, Yunran Wang, Haoran Zhang, Zhoufutu Wen, Xingwei Qu, Jian Yang, Jiaheng Liu, Minghao Liu, Xiang Yue, et al. Kor-bench: Benchmarking language models on knowledge-orthogonal reasoning tasks. *arXiv preprint arXiv:2502.04356*, 2025.

[33] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

[34] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

[35] Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. In *EMNLP*, 2023.

[36] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[37] Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. Helmet: How to evaluate long-context language models effectively and thoroughly. *arXiv preprint arXiv:2410.02694*, 2024.

[38] Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. Ruler: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*, 2024.

[39] Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. A benchmark for learning to translate a new language from one grammar book. *arXiv preprint arXiv:2402.13846*, 2024.

[40] Patrick Huber, Ernie Chang, Wei Wen, Igor Fedorov, Tarek Elgamal, Hanxian Huang, Naveen Suda, Chinnadhurai Sankar, Vish Vogeti, Yanghan Wang, et al. MobileLLM-Pro Technical Report. *arXiv preprint arXiv:2511.06719*, 2025.

[41] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015.

[42] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. TinyBERT: Distilling BERT for natural language understanding. In *Findings of EMNLP*, pages 4163–4174, 2020.

[43] Ernie Chang, Yang Li, Patrick Huber, Vish Vogeti, David Kant, Yangyang Shi, and Vikas Chandra. SDM: Scalable data mixer via influence sketching. *arXiv preprint arXiv:2505.00000*, 2025.

[44] Ernie Chang, Yang Li, Patrick Huber, Vish Vogeti, David Kant, Yangyang Shi, and Vikas Chandra. AutoMixer: Checkpoint artifacts as automatic data mixers. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 19942–19953, 2025.

[45] Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro von Werra, Thomas Wolf, Guilherme Penedo, and Hynek Kydlíček. The FineWeb datasets: Decanting the web for the finest text data at scale. *arXiv preprint arXiv:2406.17557*, 2024.

[46] Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, et al. StarCoder: May the source be with you! *arXiv preprint arXiv:2305.06161*, 2023.

[47] Keiran Paster, Marco Dos Santos, Zhangir Dongak, and Jimmy Ba. OpenWebMath: An open dataset of high-quality mathematical web text. *arXiv preprint arXiv:2310.06786*, 2023.

[48] Together Computer. RedPajama: An open source recipe to reproduce LLaMA training dataset. https://github.com/togethercomputer/RedPajama-Data, 2023.

[49] Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen McAleer, Albert Q Jiang, Jia Deng, Stella Biderman, and Sean Welleck. Llemma: An open language model for mathematics. *arXiv preprint arXiv:2310.10631*, 2023.

[50] Tri Dao and Albert Gu. Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality. *arXiv preprint arXiv:2405.21060*, 2024.

[51] Krishna C Puvvada, Faisal Ladhak, Santiago Akle Serano, Cheng-Ping Hsieh, Shantanu Acharya, Somshubra Majumdar, Fei Jia, Samuel Kriman, Simeng Sun, Dima Rekesh, and Boris Ginsburg. SWAN: An Efficient and Scalable Approach for Long-Context Language Modeling. In *Proceedings of EMNLP*, pages 2424–2438, 2025.

[52] NVIDIA Corporation. NVIDIA Blackwell Ultra Datasheet. https://resources.nvidia.com/en-us-blackwell-architecture/blackwell-ultra-datasheet, 2025.

[53] Mengzhao Chen, Chaoyi Zhang, Jing Liu, Yutao Zeng, Zeyue Xue, Zhiheng Liu, Yunshui Li, Jin Ma, Jie Huang, Xun Zhou, and Ping Luo. Scaling Law for Quantization-Aware Training. *arXiv preprint arXiv:2505.14302*, 2025.

[54] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*, 2025.

[55] 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.

[56] An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. Qwen2.5-Math Technical Report: Toward Mathematical Expert Model via Self-Improvement. *arXiv preprint arXiv:2409.12122*, 2024.

[57] Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. Qwen2.5-Coder Technical Report. *arXiv preprint arXiv:2409.12186*, 2024.

[58] Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy S Liang, Quoc V Le, Tengyu Ma, and Adams Wei Yu. DoReMi: Optimizing Data Mixtures Speeds Up Language Model Pretraining. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 69798–69818, 2023.

[59] Simin Fan, Matteo Pagliardini, and Martin Jaggi. DoGE: Domain Reweighting with Generalization Estimation. *arXiv preprint arXiv:2310.15393*, 2023.

[60] Qian Liu, Xiaosen Zheng, Niklas Muennighoff, Guangtao Zeng, Longxu Dou, Tianyu Pang, Jing Jiang, and Min Lin. RegMix: Data Mixture as Regression for Language Model Pretraining. *arXiv preprint arXiv:2407.01492*, 2024.

[61] Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, et al. Effective Long-Context Scaling of Foundation Models. *arXiv preprint arXiv:2309.16039*, 2023.

[62] Chenxin An, Fei Huang, Jun Zhang, Shansan Gong, Xipeng Qiu, Chang Zhou, and Lingpeng Kong. Training-Free Long-Context Scaling of Large Language Models. *arXiv preprint arXiv:2402.17463*, 2024.

[63] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen Technical Report. *arXiv preprint arXiv:2309.16609*, 2023.

[64] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 1877–1901, 2020.

[65] Changhan Wang, Kyunghyun Cho, and Jiatao Gu. Neural machine translation with byte-level subwords. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9154–9160, 2020.

[66] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1715–1725, 2016.

[67] Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, et al. Are We Done with MMLU? *arXiv preprint arXiv:2406.04127*, 2024.

[68] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark. *arXiv preprint arXiv:2406.01574*, 2024.

[69] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, 2023.

[70] Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, King Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, et al. SuperGPQA: Scaling LLM Evaluation Across 285 Graduate Disciplines. *arXiv preprint arXiv:2502.14739*, 2025.

[71] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

[72] 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.

[73] Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by ChatGPT really correct? rigorous evaluation of large language models for code generation. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[74] Federico Cassano, John Gouwar, Daniel Nguyen, Sydney Nguyen, Luna Phipps-Costin, Donald Pinckney, Ming-Ho Yee, Yangtian Zi, Carolyn Jane Anderson, Molly Q Feldman, et al. MultiPL-E: A Scalable and Polyglot Approach to Benchmarking Neural Code Generation. *IEEE Transactions on Software Engineering*, 49(7):3675–3691, 2023.

[75] Jacob Austin, Augustus Odena, Maxwell I Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie J Cai, Michael Terry, Quoc V Le, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

[76] Alex Gu, Baptiste Rozière, Hugh Leather, Armando Solar-Lezama, Gabriel Synnaeve, and Sida I Wang. CRUXEval: A Benchmark for Code Reasoning, Understanding and Execution. *arXiv preprint arXiv:2401.03065*, 2024.

[77] Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. Language models are multilingual chain-of-thought reasoners. In *International Conference on Learning Representations (ICLR)*, 2023.

[78] OpenAI. Multilingual Massive Multitask Language Understanding (MMMLU). https://huggingface.co/datasets/openai/MMMLU, 2024.

[79] Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Mohamed A Haggag, Snegha A, et al. INCLUDE: Evaluating Multilingual Language Understanding with Regional Knowledge. *arXiv preprint arXiv:2411.19799*, 2024.

[80] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. DeepSeek-V3 Technical Report. *arXiv preprint arXiv:2412.19437*, 2024.

[81] OLMo Team, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Costa Huang, Matt Jordan, et al. 2 OLMo 2 Furious. *arXiv preprint arXiv:2501.00656*, 2024.

[82] Janek Bevendorff, Benno Stein, Matthias Hagen, and Martin Potthast. Elastic ChatNoir: Search engine for the ClueWeb and the Common Crawl. In *European Conference on Information Retrieval (ECIR)*, pages 782–789, 2018.

[83] Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Yitzhak Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, et al. DataComp-LM: In search of the next generation of training sets for language models. *arXiv preprint arXiv:2406.11794*, 2024.

[84] Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[85] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*, 2023.

[86] Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, et al. MADLAD-400: A Multilingual And Document-Level Large Audited Dataset. *arXiv preprint arXiv:2309.04662*, 2023.

[87] Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. *arXiv preprint arXiv:2107.06499*, 2022.

[88] Alex Fang, Hadi Pouransari, Matt Jordan, Alexander Toshev, Vaishaal Shankar, Ludwig Schmidt, and Tom Gunter. Datasets, documents, and repetitions: The practicalities of unequal data quality. *arXiv preprint arXiv:2503.07879*, 2025.

[89] Niklas Muennighoff, Alexander M Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. Scaling data-constrained language models. *arXiv preprint arXiv:2305.16264*, 2025.

[90] Alexander Wettig, Kyle Lo, Sewon Min, Hannaneh Hajishirzi, Danqi Chen, and Luca Soldaini. Organize the web: Constructing domains enhances pre-training data curation. *arXiv preprint arXiv:2502.10341*, 2025.

[91] Teknium. OpenHermes 2.5: An open dataset of synthetic data for generalist LLM assistants. https://huggingface.co/datasets/teknium/OpenHermes-2.5, 2023.

[92] Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. ELI5: Long Form Question Answering. In *Proceedings of ACL*, pages 3558–3567, 2019.

[93] Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing Chat Language Models by Scaling High-Quality Instructional Conversations. In *Proceedings of EMNLP*, pages 3029–3051, 2023.

[94] Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. WildChat: 1M ChatGPT Interaction Logs in the Wild. *arXiv preprint arXiv:2405.01470*, 2024.

[95] Luca Soldaini and Kyle Lo. peS2o (Pretraining Efficiently on S2ORC) Dataset. https://github.com/allenai/pes2o, 2023.

[96] Jake Poznanski, Aman Rangapur, Jon Borchardt, Jerry Dunkelberger, Regan Huff, David Lin, Christopher Wilhelm, Kyle Lo, and Luca Soldaini. olmOCR: Unlocking trillions of tokens in PDFs with vision language models. *arXiv preprint arXiv:2502.18443*, 2025.

[97] Jiasheng Ye, Peiyu Liu, Tianxiang Sun, Junxian Zhan, Yunhua Zhou, and Xipeng Qiu. Data mixing laws: Optimizing data mixtures by predicting language modeling performance. *arXiv preprint arXiv:2403.16952*, 2025.

[98] Shizhe Diao, Yifan Yang, Yonggan Fu, Xin Dong, Dan Su, Markus Kliegl, et al. CLIMB: Clustering-based iterative data mixture bootstrapping for language model pre-training. *arXiv preprint arXiv:2504.13161*, 2025.

[99] Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, et al. SmolLM2: When smol goes big – data-centric training of a small language model. *arXiv preprint arXiv:2502.02737*, 2025.

[100] Anton Lozhkov, Raymond Li, Loubna Ben Allal, Federico Cassano, Joel Lamy-Poirier, Nouamane Tazi, et al. StarCoder 2 and The Stack v2: The Next Generation. *arXiv preprint arXiv:2402.19173*, 2024.

[101] Yixuan Li, Yaru Ma, Siyuan Yan, Chen Zhang, Junhao Liu, Jian Lu, et al. Model merging in pre-training of large language models. *arXiv preprint arXiv:2505.12082*, 2025.

[102] Zimu Lu et al. On-policy distillation: Direct alignment from demonstration to execution. *arXiv preprint arXiv:2501.00000*, 2025.

[103] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[104] W. Du, S. Toshniwal, B. Kisacanin, S. Mahdavi, I. Moshkov, G. Armstrong, S. Ge, E. Minasyan, F. Chen, and I. Gitman. Nemotron-math: Efficient long-context distillation of mathematical reasoning from multi-mode supervision. *arXiv preprint arXiv:2512.15489*, 2025.

[105] I. Moshkov, D. Hanley, I. Sorokin, S. Toshniwal, C. Henkel, B. Schifferer, W. Du, and I. Gitman. Aimo-2 winning solution: Building state-of-the-art mathematical reasoning models with openmathreasoning dataset. *arXiv preprint arXiv:2504.16891*, 2025.

[106] OpenAI. Introducing gpt 5.2, 2025.

[107] Google DeepMind. Gemini 3 pro model card, 2025.

[108] R. Li, J. Fu, B.-W. Zhang, T. Huang, Z. Sun, C. Lyu, G. Liu, Z. Jin, and G. Li. Taco: Topics in algorithmic code generation dataset. *arXiv preprint arXiv:2312.14852*, 2023.

[109] Prime Intellect. Synthetic-2 release: Four million collaboratively generated reasoning traces. Blog post, 2025.

[110] X. Zhao, Y. Liu, K. Xu, J. Guo, Z. Wang, Y. Sun, X. Kong, Q. Cao, L. Jiang, Z. Wen, et al. Small leak can sink a great ship–boost rl training on moe with icepop! *arXiv preprint arXiv:2509.00000*, 2025.

[111] L. Zhang, S. He, C. Zhang, Y. Kang, B. Li, C. Xie, J. Wang, M. Wang, Y. Huang, S. Fu, et al. Swe-bench goes live! *arXiv preprint arXiv:2505.23419*, 2025.

[112] Harbor Framework Team. Harbor: A framework for evaluating and optimizing agents and models in container environments, 2026.

[113] Y. Tian, C. Wang, Z. Liu, H. Huang, W. Yu, D. Song, J. Tang, and Y. Guo. Beyond literal mapping: Benchmarking and improving non-literal translation evaluation. *arXiv preprint arXiv:2602.00000*, 2026.

[114] Kimi Team. Kimi k2.5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026.

[115] Bita Darvish Rouhani et al. Microscaling data formats for deep learning. *arXiv preprint arXiv:2310.10537*, 2023.

[116] Moonshot AI. Kimi cli. https://www.kimi.com/code, 2026.

[117] Anthropic. Claude code. https://docs.anthropic.com/en/docs/claude-code, 2026.

[118] OpenAI. Codex. https://github.com/openai/codex, 2026.

[119] OpenClaw. Openclaw. https://docs.openclaw.ai/, 2026.

[120] Nous Research. Hermes agent. https://hermes-agent.nousresearch.com/docs/, 2026.

[121] Songlin Yang and Yu Zhang. Fla: A triton-based library for hardware-efficient implementations of linear attention mechanism. https://github.com/fla-org/flash-linear-attention, 2024.

[122] Philippe Tillet, Hsiang-Tsung Kung, and David Cox. Triton: An intermediate language and compiler for tiled neural network computations. In *Proceedings of the 3rd ACM SIGPLAN International Workshop on Machine Learning and Programming Languages (MAPL)*, 2019.

[123] Benjamin F. Spector et al. Thunderkittens: Simple, fast, and adorable kernels. In *The Thirteenth International Conference on Learning Representations*, 2025.

[124] Lei Wang et al. Tilelang: A composable tiled programming model for ai systems. *arXiv preprint arXiv:2504.17577*, 2025.

[125] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

[126] MistralAI. Mistral medium 3. https://mistral.ai/news/mistral-medium-3, 2025.

[127] Hugging Face. Open r1: A fully open reproduction of deepseek-r1. https://github.com/huggingface/open-r1, 2025.

[128] Guilherme Penedo, Anton Lozhkov, Hynek Kydlíček, Loubna Ben Allal, et al. Codeforces cots dataset. https://huggingface.co/datasets/open-r1/codeforces-cots, 2025.

[129] 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.

[130] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *Proceedings of CVPR*, 2024.

[131] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *Proceedings of ACL*, 2025.

[132] 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.

[133] 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.

[134] 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.

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

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

[137] 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.

[138] 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.

[139] 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.

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

[141] 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.

[142] 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.

[143] 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.

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

[145] 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.

[146] 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.

[147] 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.

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

[149] 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.

[150] 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.

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

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

[153] 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.

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

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

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

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

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

[159] Bruno Scherrer. Approximate policy iteration schemes: A comparison. In *International Conference on Machine Learning (ICML)*, pages 1314–1322, 2014.

[160] 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.

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

[162] 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.

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

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

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

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

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

[168] 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.

[169] 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.

[170] 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.

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

[172] 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.

[173] 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.

[174] 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.

[175] 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.

[176] 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.

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

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

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

[180] MiniMax, Aonian Li, Bangwei Gong, Bo Yang, Boji Shan, Chang Liu, Cheng Zhu, et al. Minimax-m1: Scaling test-time compute efficiently with lightning attention. *arXiv preprint arXiv:2506.13585*, 2025.

[181] 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.

[182] 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.

[183] 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.

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

[185] 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.

[186] 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.

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

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

[189] 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.

[190] 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.

[191] 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.

[192] 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.

[193] 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.

[194] 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.

[195] 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.

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

[197] 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.

[198] 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.

[199] 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.

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

[201] 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.

[202] 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.

[203] 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.

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

[205] 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.

[206] 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 *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[207] 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.

[208] 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.

[209] 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.

[210] 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, 2020.

[211] 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.

[212] 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.

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

[214] 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.

[215] 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.

[216] 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.

[217] 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.

[218] 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.

[219] 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.

[220] 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.

[221] 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.

[222] 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.

[223] 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.

[224] 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.

[225] 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.

[226] 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)*, pages 35–50, 2014.

[227] NVIDIA. NVIDIA Nemotron 3: Efficient and Open Intelligence. *arXiv preprint arXiv:2512.20856*, 2025.

[228] L. Wang, Y. Cheng, Y. Shi, Z. Mo, et al. Tilelang: Bridge programmability and performance in modern neural kernels. *International Conference on Learning Representations (ICLR)*, 2026.

[229] L. Haas, G. Yona, G. D'Antonio, S. Goldshtein, and D. Das. Simpleqa verified: A reliable factuality benchmark to measure parametric knowledge. *arXiv preprint arXiv:2509.07968*, 2025.

[230] Y. He, S. Li, J. Liu, Y. Tan, et al. Chinese simpleqa: A chinese factuality evaluation for large language models. *arXiv preprint arXiv:2411.07140*, 2024.

[231] L. Phan, A. Gatti, Z. Han, N. Li, et al. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2025.

[232] G. Tsoukalas, J. Lee, J. Jennings, J. Xin, et al. Putnambench: Evaluating neural theorem-provers on the putnam mathematical competition. *arXiv preprint arXiv:2407.11214*, 2024.

[233] OpenAI. Openai mrcr: Long context multiple needle in a haystack benchmark, 2024.

[234] Z. Lu, C. Li, Y. Shi, W. Shen, et al. Corpusqa: A 10 million token benchmark for corpus-level analysis and reasoning. *arXiv preprint arXiv:2601.14952*, 2026.

[235] M. A. Merrill, A. G. Shaw, N. Carlini, B. Li, et al. Terminal-bench: Benchmarking agents on hard, realistic tasks in command line interfaces. *arXiv preprint arXiv:2601.11868*, 2026.

[236] OpenAI. Introducing swe-bench verified, 2024.