每日论文分析 (2026-02-18)

标签: [强推], [RLVR], [训练稳定性], [字节], [Google], [Cornell]

论文1: STAPO – Stabilizing Reinforcement Learning for LLMs by Silencing Rare Spurious Tokens

基本信息

- **标题**: STAPO: Stabilizing Reinforcement Learning for LLMs by Silencing Rare Spurious Tokens
- **ArXiv ID**: 2602.15620
- **作者**: Shiqi Liu, Zeyu He, Guojian Zhan, Letian Tao, Zhilong Zheng, Jiang Wu, Yinuo Wang, Yang Guan, Kehua Sheng, Bo Zhang, Keqiang Li, Jingliang Duan, Shengbo Eben Li
- **机构**: 清华大学、 Cornell University
- **提交日期**: 2026-02-17

动机

问题形式化: 这篇论文解决的是 RL 微调大语言模型时的训练不稳定性问题。现有 RL 方法（如 GRPO、DPO）依赖熵正则化等启发式技术来维持稳定性，但仍常遭遇**后期性能崩溃**（late-stage performance collapse）。

数学框架: 作者推导出 token 级别策略梯度与 token 概率和局部策略熵的负相关关系：

$$\|\nabla_\theta \mathcal{L}_{RL}\| \propto \frac{1}{p(t) \cdot H(t)}$$

其中 $p(t)$ 是 token t 的概率，$H(t)$ 是局部策略熵。这意味着低概率 token 会获得异常大的梯度。

核心发现: 训练不稳定性由约 0.01% 的极少量 token 驱动，作者称之为 **spurious tokens**（伪tokens）。这些 tokens 出现在正确响应中时，对推理结果贡献极小，却继承完整序列级奖励，导致梯度异常放大。

公式讲解

核心公式是作者对梯度幅度的推导：

$$\|\nabla_\theta \mathcal{L}_{token}\| \approx R \cdot \frac{1 - \pi_\theta(t|c)}{Z} \cdot \nabla_\theta \log \pi_\theta(t|c)$$

- R: 序列级奖励
- $\pi_\theta(t|c)$: 条件概率
- Z: 归一化因子

关键洞察: 当 $\pi_\theta(t|c) \to 0$ 时，梯度趋向无穷大。Spurious tokens 的概率极低但获得高奖励，导致梯度爆炸。

STAPO 解决方案:

$$\mathcal{L}_{STAPO} = \frac{1}{|\mathcal{T}_{valid}|} \sum_{t \in \mathcal{T}_{valid}} \log \pi_\theta(t|c) \cdot R$$

通过 mask 掉 spurious tokens 并在有效 tokens 上重新归一化损失。

实验设计

- **基线方法**: GRPO, 20-Entropy, JustRL
- **数据集**: 6 个数学推理基准（MATH, GSM8K 等）
- **模型**: Qwen 1.7B, 8B, 14B
- **评估指标**: 准确率、熵稳定性

实验结论

1. **主实验结果**:
   - STAPO 相比 GRPO 平均提升 **7.13%**
   - 在所有模型规模上保持一致的熵稳定性
2. **消融实验**:
   - 验证了 spurious tokens 确实约占比 0.01%
   - Mask 策略比 reweighting 更有效

严厉审视

优点: – 理论推导严谨，从梯度幅度公式出发 – 识别了一个关键的训练不稳定性来源 – 方法直接有效

潜在问题: – 0.01% 的阈值如何自适应确定？论文未讨论 – 仅在数学推理任务上验证，泛化性待检验 –“spurious”的定义依赖具体任务，可能存在任务依赖性

论文2: On Surprising Effectiveness of Masking Updates in Adaptive Optimizers

基本信息

- **标题**: On Surprising Effectiveness of Masking Updates in Adaptive Optimizers
- **ArXiv ID**: 2602.15322
- **作者**: Taejong Joo, Wenhan Xia, Cheolmin Kim, Ming Zhang, Eugene Ie
- **机构**: Google
- **提交日期**: 2026-02-17

动机

问题: LLM 训练几乎 exclusively 使用密集自适应优化器（Adam 系列），但作者发现**随机 mask 参数更新**反而更有效。

核心发现: 随机 masking 引入了一种**曲率依赖的几何正则化**，平滑了优化轨迹。

公式讲解

Magma (Momentum-aligned gradient masking):

$$g_{magma} = m \odot g + (1 - m) \odot (g \cdot \frac{m^\top v}{\|m\|^2})$$

- $m \in \{0, 1\}^d$: 二值 mask 向量
- v: 动量向量
- g: 原始梯度

物理意义: 这实际上是将被 mask 的梯度方向与动量对齐，形成一种”软约束”，使得 masked updates 不会偏离动量方向太远。

实验设计

- **基线**: Adam, Muon, Sophia, AdamW
- **模型规模**: 1B 参数
- **任务**: 语言模型预训练
- **评估指标**: Perplexity

实验结论

- Magma 相比 Adam 降低 perplexity **19%**
- 相比 Muon 降低 **9%**
- 计算开销可忽略

严厉审视

优点: – 简单有效的drop-in替换 – 理论分析（几何正则化）有洞见

问题: – 未在大规模（>10B）模型上验证 – 1B 规模的优势是否能保持是未知数 – 与现有 LR 调度器的交互未充分讨论

论文3: GLM-5 – from Vibe Coding to Agentic Engineering

基本信息

- **标题**: GLM-5: from Vibe Coding to Agentic Engineering
- **ArXiv ID**: 2602.15763
- **作者**: GLM-5 Team (186位作者)
- **机构**: 智谱AI (Zhipu AI), 清华大学
- **提交日期**: 2026-02-17

动机

核心贡献: 1. **DSA (Deferred Speculative Approximation)**: 减少训练和推理成本，同时保持长上下文保真度 2. **异步 RL 基础设施**: 解耦生成与训练，提高后训练效率 3. **异步 Agent RL 算法**: 从复杂、长程交互中学习

技术亮点

DSA: 这是一种推测解码/训练加速技术，类似于 Medusa、Speculative Decoding，但应用于训练阶段。

异步 RL: 传统 RL 需要同步生成和训练，智谱提出解耦架构： – 离线生成 responses – 批量训练 – 异步流水线提高吞吐量

实验结论

- 在主流 benchmark 上达到 SOTA
- 在真实软件工程任务上表现突出
- 端到端编码能力显著提升

严厉审视

优点: – 工业级系统工作，186人团队 – 异步 RL 基础设施有工程价值 – Agent 能力展示有说服力

问题: – 论文细节有限（arXiv 仅有 4 页） – 很多技术细节未公开 – “SOTA” 需要具体 benchmark 数据支撑

对比与总结

论文	方向	核心贡献	评价
STAPO	RL训练稳定性	识别并解决 spurious tokens	⭐⭐⭐⭐⭐ 理论扎实
Magma	优化器	梯度masking + 动量对齐	⭐⭐⭐⭐ 简单有效
GLM-5	Agent系统	异步RL + DSA	⭐⭐⭐⭐ 工程强大

今日推荐: STAPO 是最具理论价值的工作，对 GRPO/DPO/RLVR 路线有直接启示。Magma 的几何正则化视角有趣。GLM-5 展示了工程能力但细节不足。