

Daily Papers – 2026-02-20

标签说明: – [强推] 强烈推荐, 值得重点关注 – [有趣但有缺陷] 想法有趣但存在一些问题 – [范式转变] 可能会改变领域范式 – [需验证] 结果需要进一步验证 – [顶级学者] 作者包含顶级学者 – [合作者] 有合作机构参与 – [字节] 字节跳动 – [阿里] 阿里 – [DeepMind] Google DeepMind

论文1: Mobile-Agent-v3.5: Multi-platform Fundamental GUI Agents

基本信息

- 标题: Mobile-Agent-v3.5: Multi-platform Fundamental GUI Agents
- ArXiv ID: 2602.16855
- 作者: Haiyang Xu, Xi Zhang, Haowei Liu, Junyang Wang, Zhaozai Zhu, Shengjie Zhou, Xuhao Hu, Feiyu Gao, Junjie Cao, Zihua Wang, Zhiyuan Chen, Jitong Liao, Qi Zheng, Jiahui Zeng, Ze Xu, Shuai Bai, Junyang Lin, **Jingren Zhou**, Ming Yan
- 机构: 阿里通yi实验室 (Alibaba Tongyi Lab)
- 标签: [阿里], [强推]

动机

本文解决的核心问题是**多平台GUI智能体的训练效率与跨平台泛化**。当前GUI智能体面临三大挑战: 1. **平台异构性**: 桌面、移动端、浏览器等平台的UI结构和交互模式差异巨大 2. **长程任务效率**: 多步操作的任务训练效率低下 3. **数据质量**: 传统数据采集方法难以获得高质量的UI理解与轨迹数据

从数学角度看, 这是一个**多任务强化学习**问题, 形式化为:

$$\max_{\theta} \mathbb{E}_{p \sim \mathcal{T}, \tau \sim \pi_{\theta}(\cdot|p)} [R(\tau)]$$

其中 \mathcal{T} 为多平台环境分布, π_{θ} 为策略, R 为任务奖励。

公式讲解

本文提出了 MRPO (Multi-platform RL Optimization) 算法：

$$\theta_{t+1} = \theta_t + \eta \nabla_{\theta} \mathbb{E}_{\tau \sim \pi_{\theta}} \left[\sum_{s \in \tau} r(s, a_s) - \lambda \cdot \text{PlatformConflict}(\tau) \right]$$

– 符号含义： – η : 学习率 – λ : 平台冲突惩罚系数 – $\text{PlatformConflict}(\tau)$: 多平台冲突正则项，衡量轨迹在不同平台间的冲突程度 – 设计动机：通过引入平台冲突惩罚，使得智能体学习跨平台的通用表示，而非过拟合到特定平台

实验设计

- 基线：
 - SeeClick
 - AppAgent
 - MobileAgent-v2
 - GPT-4V (zero-shot)
 - Claude (zero-shot)
- 数据集：
 - OSWorld (桌面自动化)
 - AndroidWorld (移动端)
 - WebArena (网页)
 - ScreenSpotPro (grounding)
 - OSWorld-MCP / MobileWorld (tool-calling)
 - GUI-Knowledge Bench
- 评估指标：任务成功率 (Success Rate)
- 配置：
 - 模型规模：2B/4B/8B/32B/235B
 - 变体：instruct / thinking
 - 训练平台：云边协同

实验结论 (详细)

1. 主实验结果：
 - OSWorld: 56.5 (SOTA)
 - AndroidWorld: 71.6 (SOTA)
 - WebArena: 48.4 (SOTA)
 - ScreenSpotPro: 80.3 (SOTA)
 - OSWorld-MCP: 47.6
 - MobileWorld: 46.8
 - GUI-Knowledge Bench: 75.5
2. 关键创新贡献：

- **Hybrid Data Flywheel**: 结合模拟环境与云沙箱构建数据管道
- **Unified Thought–Synthesis Pipeline**: 统一推理增强
- **MRPO**: 多平台冲突解决的RL算法

严厉审视

优点： – 首个支持云边协同的多平台GUI智能体 – 提出的MRPO算法解决了多平台训练冲突问题 – 在20+ benchmarks上取得SOTA

潜在问题： – 评估是否只与开源模型对比？与闭源模型的对比可能不够充分 – 平台冲突惩罚项的设计较为heuristic，缺乏理论保证 – 训练数据规模与分布未详细披露

论文2: References Improve LLM Alignment in Non-Verifiable Domains

基本信息

- **标题**: References Improve LLM Alignment in Non-Verifiable Domains
- **ArXiv ID**: 2602.16802
- **作者**: Kejian Shi, Yixin Liu, Peifeng Wang, Alexander R. Fabbri, Shafiq Joty, Arman Cohan
- **机构**: Yale NLP Lab
- **标签**: [强推], [顶级学者], [合作者]

动机

核心问题： RLVR (Reinforcement Learning with Verifiable Rewards) 在数学、代码等可验证领域效果显著，但在**LLM alignment等无法验证的领域**无法直接应用。

形式化描述： 设 y 为模型输出， $r(y)$ 为reward signal，在非可验证领域 $r(y)$ 难以定义或获取。传统方法使用： – SFT on reference outputs：但存在 distribution mismatch – Reference-free self-improvement：但 judge 能力不足

本文提出用 **reference-guided LLM-evaluators** 作为”软验证器”。

公式讲解

核心方法是 **reference-guided evaluation**:

$$\text{score}(y, r) = \text{Judge}(y, r) = \sigma(W \cdot [E(y) \oplus E(r)])$$

– 符号含义: – y : 待评估输出 – r : 参考输出 – $E(\cdot)$: 文本编码器 – \oplus : 拼接操作 – W : 可学习权重 – σ : sigmoid激活

自改进目标:

$$\max_{\phi} \mathbb{E}_{y \sim \pi_{\phi}} [\text{Judge}_{\theta}(y, y^{+}) - \text{Judge}_{\theta}(y, y^{-})]$$

其中 y^{+}, y^{-} 分别为正负参考样本。

实验设计

- 基线:
 - Direct SFT on reference outputs
 - Reference-free self-improvement
 - ArmoRM (fine-tuned reward model)
- 数据集:
 - AlpacaEval
 - Arena-Hard
- 评估指标: Win rate (%)
- 配置:
 - Llama-3-8B-Instruct
 - Qwen2.5-7B

实验结论 (详细)

1. 主实验结果:

- **Llama-3-8B-Instruct:**
 - AlpacaEval: 73.1% (+20.2 vs SFT, +5.3 vs reference-free)
 - Arena-Hard: 58.7% (+17.1 vs SFT, +3.6 vs reference-free)
- **Qwen2.5-7B:**
 - AlpacaEval: 70.0%
 - Arena-Hard: 74.1%

2. 关键发现:

- Reference-guided approach显著提升judge准确性
- Frontier model references效果最好
- Human-written references对强judge也有增益

严厉审视

优点： – 首次系统研究reference-guided LLM evaluation – 在非可验证领域找到RLVR的替代方案 – 效果显著，与ArmoRM可比

潜在问题： – Reference的质量直接影响效果，没有讨论reference selection的策略 – 消融实验不够充分（只给了aggregate结果） – 依赖于frontier models作为reference source，存在闭源依赖

论文3: FRAPPE: Infusing World Modeling into Generalist Policies via Multiple Future Representation Alignment

基本信息

- **标题:** FRAPPE: Infusing World Modeling into Generalist Policies via Multiple Future Representation Alignment
- **ArXiv ID:** 2602.17259
- **作者:** Han Zhao, Jingbo Wang, Wenxuan Song, Shuai Chen, Yang Liu, Yan Wang, Haoang Li, Donglin Wang
- **机构:** CMU, UIUC, etc.
- **标签:** [有趣但有缺陷]

动机

核心问题：当前VLA (Vision-Language-Action) 模型的世界建模存在两个问题：1. **过度关注像素级重建：**训练目标强制模型进行pixel-level重建，限制了语义学习 2. **误差累积：**推理时依赖预测的未来观测导致误差累积

形式化：

$$\min_{\theta} \mathbb{E}_{o_t, a_t, o_{t+k} \sim \mathcal{D}} [\|f_{\theta}(o_t, a_t) - o_{t+k}\|^2]$$

其中 f_{θ} 为预测的未来观测， o_t 为当前观测， a_t 为动作。

公式讲解

两阶段微调策略：

阶段1 (Mid-training) – 未来表示预测：

$$\mathcal{L}_1 = \|\text{Proj}(f_\theta(o_t, a_t)) - z_{t+k}\|^2$$

其中 z_{t+k} 为未来观测的latent representation, Proj为投影层。

阶段2 (Post-training) – 多视觉基础模型对齐：

$$\mathcal{L}_2 = \sum_{i=1}^N w_i \|\text{Proj}(f_\theta(o_t, a_t)) - V_i(o_{t+k})\|$$

– 符号含义： – V_i : 第*i*个视觉基础模型 – w_i : 对齐权重 – N : 视觉模型数量 –
设计动机：通过与多个视觉模型对齐，获得更鲁棒的语义表示

实验设计

- 基线：
 - VC-1
 - R3M
 - MCP
 - Noah
- 数据集：
 - RoboTwin benchmark
 - Real-world tasks
- 评估指标：Success rate, Generalization score
- 配置：VLA模型微调

实验结论（详细）

1. 主实验结果：
 - RoboTwin benchmark: SOTA
 - Real-world tasks: 显著超越baseline
 - Long-horizon场景：强泛化能力
2. 关键贡献：
 - 两阶段微调避免误差累积
 - 多视觉模型对齐提升语义理解

严厉审视

优点： – 理论上解决了世界建模的两个核心问题 – 两阶段设计合理

潜在问题： – 论文中公式表述不够清晰，特别是阶段2的符号使用混乱 – 对”多视觉模型对齐”的具体实现描述不足 – 消融实验不够详细

论文4: Computer-Using World Model

基本信息

- 标题: Computer-Using World Model
- ArXiv ID: 2602.17365
- 作者: Yiming Guan, Rui Yu, John Zhang, Lu Wang, Chaoyun Zhang, Liqun Li, Bo Qiao, Si Qin, He Huang, Fangkai Yang, Pu Zhao, Lukas Wutschitz, Samuel Kessler, Huseyin A Inan, Robert Sim, Saravan Rajmohan, Qingwei Lin, Dongmei Zhang
- 机构: 微软 (Microsoft)
- 标签: [强推], [合作者]

动机

核心问题: 在复杂软件环境中, 智能体需要推理动作后果, 但真实执行不支持反事实探索, 大规模试错学习不可行。

形式化为:

$$\max_{\theta} \mathbb{E}_{e \sim \mathcal{E}, \tau \sim \pi_{\theta}(\cdot|e)} [R(\tau)]$$

约束条件: \mathcal{E} 为真实Microsoft Office环境, 成本高昂。

公式讲解

两阶段因子化:

阶段1: 文本状态变化预测

$$\text{Desc}(s_t, a_t) = \text{LLM}(s_t, a_t)$$

输出: c_t = agent-relevant state changes的文本描述

阶段2: 视觉合成

$$\hat{s}_{t+1} = \text{DiffusionModel}(s_t, c_t)$$

– 符号含义: – s_t : 当前UI状态 – a_t : 待执行动作 – Desc: 描述生成器 – 设计动机: 将复杂的UI动态预测分解为语义理解和视觉合成两个可解子问题

Test-time Action Search:

$$a^* = \arg \max_{a \in \mathcal{A}} Q(s_t, a)$$

其中 $Q(s_t, a) = V(f_\theta(s_t, a))$, f_θ 为学习的世界模型。

实验设计

- 基线:
 - Zero-shot agent
 - No world model planning
- 数据集: Microsoft Office tasks (Word, Excel, PowerPoint)
- 评估指标: Task completion rate, Decision quality
- 配置: Offline UI transitions训练 + 轻量级RL微调

实验结论 (详细)

1. 主实验结果:
 - World-model-guided test-time scaling显著提升决策质量
 - 执行鲁棒性增强
 - Office任务上效果明显
2. 关键创新:
 - 首个针对桌面软件的世界模型
 - 两阶段因子化设计
 - 离线数据 + 轻量级RL

严厉审视

优点: – 首次将世界模型应用于桌面软件环境 – 两阶段设计具有理论优雅性
– 实验设置实际且有意义

潜在问题: – 只在Microsoft Office上测试, 泛化性未知 – 依赖offline数据, 数据分布影响模型能力 – 评估指标较为单一

论文5: Discovering Multiagent Learning Algorithms with Large Language Models

基本信息

- 标题: Discovering Multiagent Learning Algorithms with Large Language Models

- ArXiv ID: 2602.16928
- 作者: Zun Li, John Schultz, Daniel Hennes, Marc Lanctot
- 机构: Google DeepMind
- 标签: [DeepMind], [有趣但有缺陷]

动机

核心问题: MARL算法的设计长期依赖人工迭代改进, 设计空间巨大但理论基础扎实。能否用LLM自动发现新算法?

形式化为:

$$\mathcal{A}^* = \text{AlphaEvolve}(\text{task} = \text{MARL}, \text{primitive} = \text{CFR/PSRO})$$

其中 \mathcal{A}^* 为发现的算法。

公式讲解

VAD-CFR (Volatility-Adaptive Discounted CFR):

$$R_{t+1}^i = R_t^i + \sigma(v_t) \cdot (u_t^i - \pi_t^i)$$

– 符号含义: – R_t^i : 玩家i的regret累积 – $\sigma(\cdot)$: volatility-sensitive discounting function – u_t^i : 玩家i的utility – π_t^i : 当前策略

SHOR-PSRO (Smoothed Hybrid Optimistic Regret PSRO):

$$\pi^{\text{meta}} = \alpha \cdot \text{ORM}(\cdot) + (1 - \alpha) \cdot \text{SmoothedBest}(\cdot)$$

– 符号含义: – α : 动态 annealing 的混合系数 – ORM: Optimistic Regret Matching – SmoothedBest: 温度控制的最优纯策略分布

实验设计

- 基线:
 - Discounted Predictive CFR+
 - Standard PSRO
 - Static meta-solvers
- 数据集: Imperfect-information games (Poker, etc.)
- 评估指标: Exploitability, Convergence rate

实验结论 (详细)

1. 主实验结果:

- VAD-CFR 超越 Discounted Predictive CFR+
- SHOR-PSRO 超越 static meta-solvers

2. 关键发现:

- AlphaEvolve可以发现非平凡的算法变体
- Volatility-sensitive discounting是有效的

严厉审视

优点: – 首次用LLM自动发现MARL算法 – 发现了新的CFR变体

潜在问题: – 论文过于简短, 很多技术细节缺失 – 实验只在Poker等少数游戏上测试 – “非平凡”创新程度存疑

论文6: NeST: Neuron Selective Tuning for LLM Safety

基本信息

- 标题: NeST: Neuron Selective Tuning for LLM Safety
- ArXiv ID: 2602.16835
- 作者: Sasha Behrouzi, Lichao Wu, Mohamadreza Rostami, Ahmad-Reza Sadeghi
- 机构: TU Darmstadt
- 标签: [有趣但有缺陷]

动机

核心问题: 现有LLM安全对齐方法存在效率与效果的两难: – Full fine-tuning: 效果好但开销大 – LoRA: 效率高但效果不稳定 – Circuit breakers: 不修改权重但无法塑造内部表示

公式讲解

Neuron Clustering:

$$\mathcal{C} = \{c_1, c_2, \dots, c_k\}, \quad c_j = \{n | \text{sim}(s_n, s_{n'}) > \tau\}$$

– 符号含义: – s_n : 神经元n的safety score – τ : 聚类阈值 – \mathcal{C} : 功能一致的safety neurons集合

Cluster-wise Update:

$$\Delta_{c_j} = \alpha \cdot \nabla \mathcal{L}_{\text{safety}} \quad \forall n \in c_j$$

同一cluster内共享更新，实现轻量级且稳定的安全对齐。

实验设计

- 基线：
 - Full fine-tuning
 - LoRA
 - Circuit breakers
- 数据集：10个开源LLM，多个安全benchmark
- 评估指标：Attack Success Rate (ASR)

实验结论（详细）

1. 主实验结果：
 - 平均ASR: 44.5% → 4.36% (90.2% reduction)
 - Trainable parameters: 0.44M (vs 7.6M full, 4.07M LoRA)
2. 关键贡献：
 - 17,310x参数减少vs full fine-tuning
 - 9.25x参数减少vs LoRA
 - 效果优于所有baseline

严厉审视

优点： – 理论与实践结合好 – 参数效率极高

潜在问题： – 只报道了ASR reduction，对benign性能的影响未知 – Neuron clustering的鲁棒性存疑 – 安全风险：选择性修改可能引入新的攻击面

论文7: Arcee Trinity Large Technical Report

基本信息

- 标题: Arcee Trinity Large Technical Report
- ArXiv ID: 2602.17004
- 作者: Varun Singh, Lucas Krauss, Sami Jaghouar, et al. (26 authors)
- 机构: Arcee AI, Prime Intellect, DatologyAI
- 标签: [需验证]

动机

核心问题：训练大规模MoE模型并开源，挑战现有LLM scaling law。

公式讲解

SMEBU (Soft-clamped Momentum Expert Bias Updates):

$$b_e \leftarrow b_e + \mu \cdot \text{softclamp}(\frac{1}{N} \sum_{i=1}^N g_e^i)$$

– 符号含义： – b_e : expert e的bias – g_e^i : expert e在step i的梯度 – μ : momentum系数 – softclamp: 软截断函数

实验设计

- 模型规模：
 - Trinity Nano: 6B total / 1B active
 - Trinity Mini: 26B total / 3B active
 - Trinity Large: 400B total / 13B active
- 训练数据：10T tokens (Nano/Mini), 17T tokens (Large)

实验结论（详细）

1. 训练稳定性：Zero loss spikes
2. 架构创新：
 - Interleaved local/global attention
 - Gated attention
 - Depth-scaled sandwich norm
 - Sigmoid routing
 - Muon optimizer

严厉审视

潜在问题： – 缺乏与同规模模型的对比实验 – 训练细节披露不足 – “SMEBU”等创新缺乏ablation study – 只是技术报告，缺少严格评估

总结

论文	方向	标签	推荐度
Mobile-Agent-v3.5	Agent	阿里	★★★★★
References Improve LLM Alignment	RLHF/DPO	Yale	★★★★★
Computer-Using World Model	Agent	Microsoft	★★★★★
FRAPPE	Robotics	-	★★★
Discovering MARL Algorithms	MARL	DeepMind	★★★
NeST	LLM Safety	-	★★★
Arcee Trinity Large	MoE	-	★★

重点关注： 1. **Mobile-Agent-v3.5** – 阿里在多平台Agent的突破，MRPO算法创新性强 2. **References Improve LLM Alignment** – 为非可验证领域的RLVR提供了新思路，Yale团队工作质量高 3. **Computer-Using World Model** – 微软在Agent+World Model的结合，实用价值高