数学推論能力を向上させる大規模言語モデルのファインチューニングに関する研究

折池雄太（AutoRes）恩田直登（OSX）平澤寅庄（OSX）
園田翔（理研）三内顕義（京都大・NII）

はじめに

AutoRes:

- 研究の自動化を目指すプロジェクト（大学や企業の研究者、学生で構成）
- 取り組み：AI scientist的なこと。数学研究の自動化。

この分野に興味のある方
共同研究者を求めているのでお声がけください

数学研究の自動化：

- lean：数学定理を厳密に証明するのに設計されたプログラミング言語
- leanを用いてAIが高度な数学の定理を自律的に証明できるようにしたい
  →仮説生成より仮説検証（＝定理の証明能力）がボトルネックと分かった
  →reasoning能力の向上が必須で、reasoningモデルの研究を行っていた

研究目的

- 数学推論タスクにおけるLLMの性能を向上させる手法の探索
  - SFT
  - accuracy rewardに基づく強化学習
  - step-by-step PPO

関連研究(数学推論能力の向上)

CoT (Chain-of-Thought):

- 推論過程を可視化・明示することで正答率向上

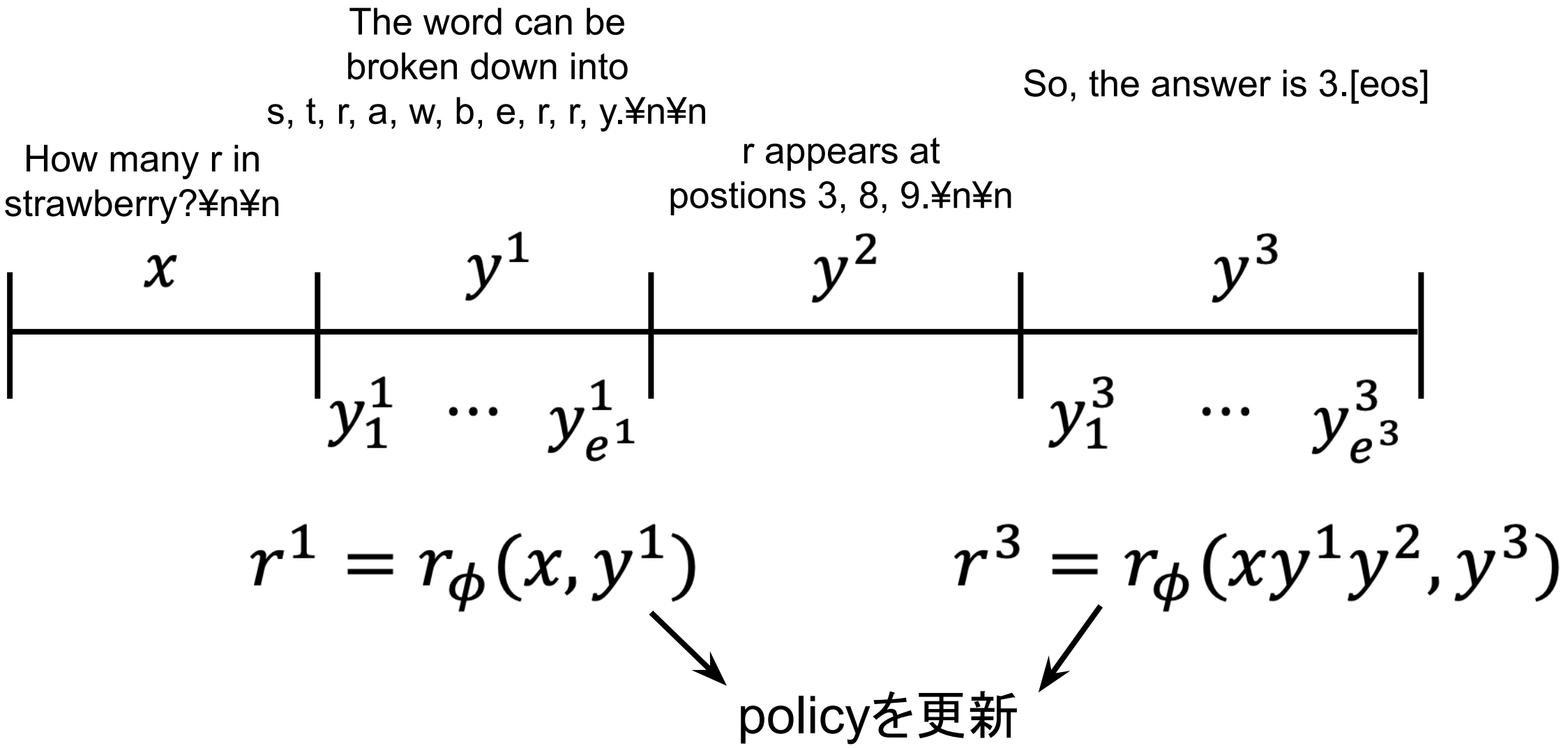
Accuracy reward:

- DeepSeek-R1: 最終解答の正誤を直接報酬とする手法

PRM (Process-Reward Model):

- 推論過程の中間ステップを評価する報酬モデルを用いて精度向上
- Math-Shepherd: step-by-step PPOの導入で正答率向上

step-by-step PPO



実験設定: データセットとモデル

使用データセット :

- NuminaMath-CoT: 数学問題と推論過程を含む大規模データセット（約86万件のデータ）

ベースモデル :

- llm-jp-3-13b: 英語・日本語のバイリンガルモデル（コンペティション指定）
- Qwen2.5-Math-7B: 数学特化の事前学習済みモデル．英語・中国語のバイリンガルモデル（比較実験用）

手法(1): Supervised Fine-Tuning (SFT)

アプローチ:

- NuminaMath-CoTデータセットを用いたSFTを2epoch実施
- 翻訳コスト・精度問題を避けるため英語データのまま llm-jp-3-13Bを訓練

ねらい:

- CoTの学習を通じた基本的な数学的推論能力の獲得
- 強化学習の前段階としての基本的な数学的推論能力の獲得

手法(2): 強化学習（accuracy reward, GRPO）

アプローチ：

- DeepSeek-R1で用いられているaccuracy reward（解答の正誤を直接報酬）に基づく強化学習
- Hugging Faceのopen-r1ライブラリを活用した実装

実験結果

正答率は事務局から提供された日本語訳した　MATHデータを使用

- **llm-jp-3-13B + NuminaMath-CoT (SFTのみ):**
  - 正答率19%

- **llm-jp-3-13B + SFT + GRPO:**
  - 正答率19%→継続学習で13%に低下

今回提出した
モデル

- **Qwen2.5-Math-7B-Instruct + GRPO :**
  - 正答率65%→70%へ上昇を維持

その他の取り組み: KLペナルティによる安定性

KLペナルティの強度 beta の違いにより、学習の安定性が異なる
ことが確認できた

- **Qwen2.5-Math-7B-Instruct + GRPO (beta=0.04):**
  - 正答率65%→70%へ上昇を維持

- **Qwen2.5-Math-7B + GRPO (beta=0):**
  - 正答率12%→28%へ上昇後、学習崩壊し0%に

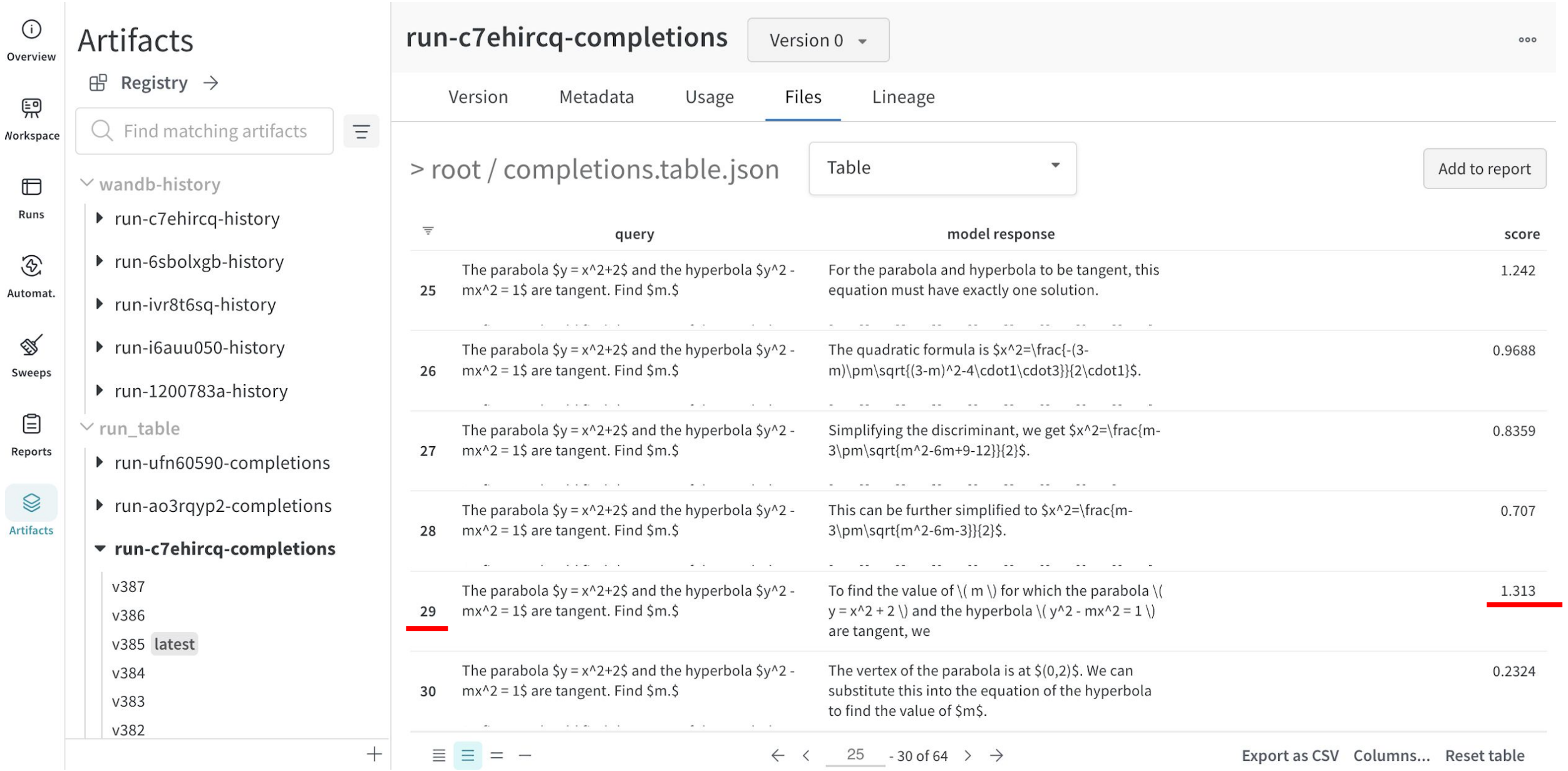
その他の取り組み: step-by-step PPO

概要:

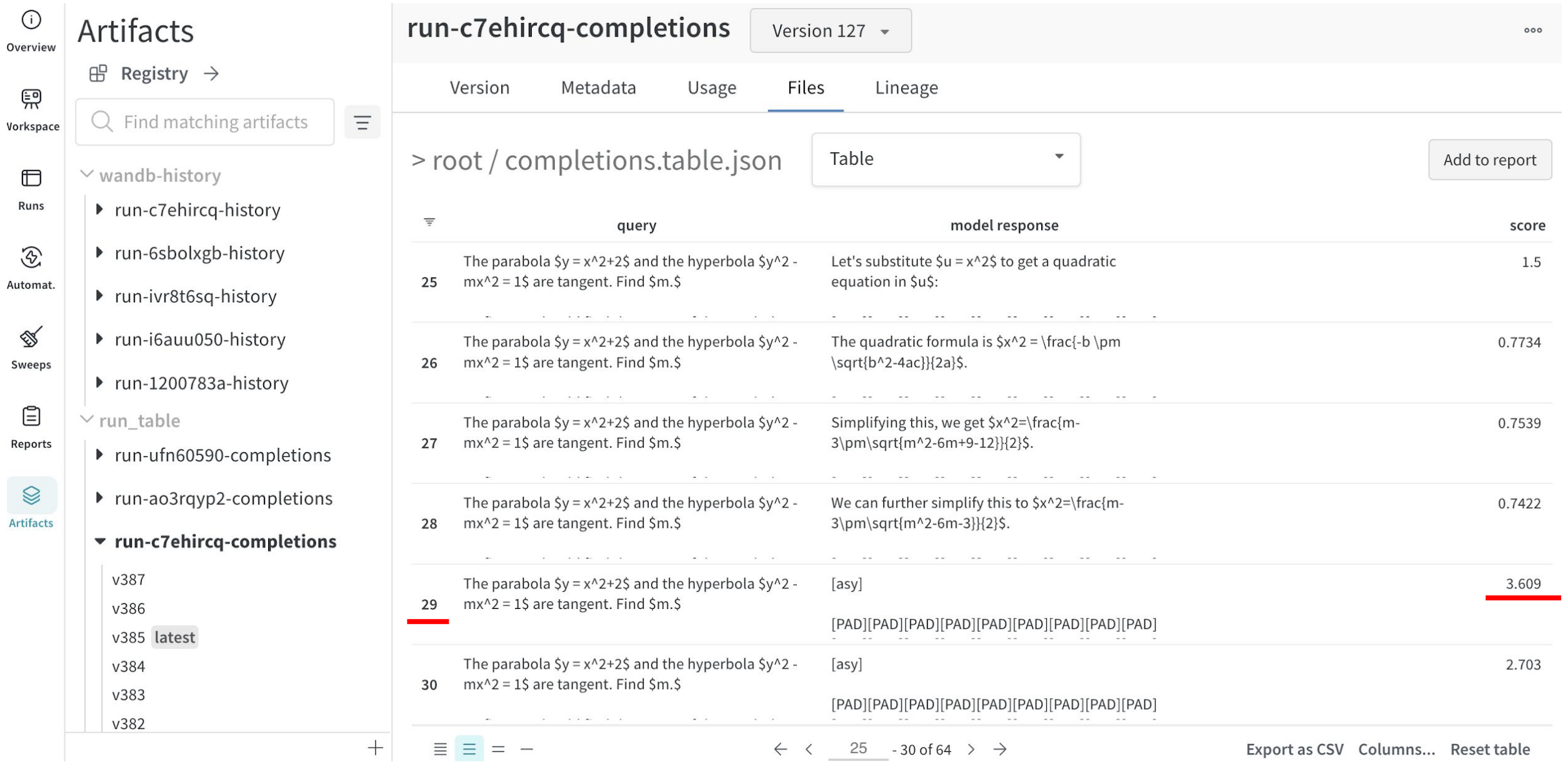
- 各推論ステップに対して個別の報酬を与えてPPOを行う手法

直面した課題:

- **報酬ハッキング:** 不自然な短文でも高報酬になる問題
  - PRMの精度が強化学習の成否を左右
  - 不適切な報酬設計が「望ましくない行動」を強化する危険性
- **トークナイザー不一致:** policyモデルとPRMの互換性問題
- **PRMの学習コスト:**
  - 数学データによる大規模事前学習
  - Process Supervision Datasetが十分な量必要
  - Process Supervision Datasetを用いたSFTの実施



強化学習が進むほどresponseが変でもscoreが高いものを見つけて最適化してしまう



強化学習が進むほどresponseが変でもscoreが高いものを見つけて最適化してしまう

結論・今後の展望

結論：

- 数学推論タスクでのファインチューニング事例を検証
- SFTでも一定の性能向上、強化学習には報酬設計と安定化が課題
- 数学事前学習の充実度が後続の強化学習効果に大きく影響

今後の展望：

- より堅牢な報酬モデル設計と報酬ハッキング対策
- 数学推論に特化した事前学習の拡充
- leanへの拡張の検討

この分野に興味のある方
共同研究者を求めているのでお声がけください