Paper Summary (Analytical)

Paper: Improving Generative Ad Text on Facebook using Reinforcement Learnin
(arXiv:2507.21983v2, Dec 15, 2025)

1. Main Problem and Why It Matters

Large language models (LLMs) are powerful, but they often need post-training
to perform well on real tasks. Most work on post-training focuses on subjective
human preferences (e.g., RLHF). This paper asks a more concrete question: Ca
reinforcement learning (RL) with real performance metrics measurably improve
high-stakes, real-world product? The domain is online advertising, where small
click-through rate (CTR) can represent large economic gains. The key contribut
quantify the impact of RL post-training using actual advertiser outcomes rathe
offline benchmarks.

2. Inputs and Outputs of the System

- Input: A human-written ad text provided by an advertiser.

- Output: Multiple AI-generated variations of the ad text, which the advertiser c
and edit before publishing.

The output is not just a single best rewrite. The system supports exploration by
multiple options so the advertiser can select the most appropriate variant.

3. Data Used (Type, Source, Size)

The method relies on historical advertising performance data from Meta's "Mul
Optimization," where advertisers test multiple text variants while keeping othe
fixed (image, targeting, etc.). This makes it possible to attribute performance
differences to the text itself.

Key properties:

- Type: Historical ad performance pairs (text variants with the same non-text at

- Source: Meta/Facebook ad platform data.

- Size: In the live A/B test, the model was evaluated over ~35,000 advertisers a
~640,000 ad variations over 10 weeks.

4. Proposed Method in Simple Terms (RLPF)

The paper introduces Reinforcement Learning with Performance Feedback (RLP

1. Reward Model Training:

- Build pairs of ad texts where one performed better (higher CTR) than the ot
multitext historical data.

- Train a reward model to predict preference: a text with higher CTR should ge
higher reward score.

2. RL Fine-Tuning with PPO:

- Use the reward model as a proxy for real-world feedback.

- Fine-tune the LLM with Proximal Policy Optimization (PPO) to maximize rewa
staying close to the base model.

- Add a length penalty to avoid overly long ad text.

This is conceptually between RLHF (human preference labels) and RL with verif
(math/coding). Here the reward is a business metric (CTR).

Simple Pseudocode

5. Models Compared

- Imitation LLM v1: SFT on synthetic ad rewrites.

- Imitation LLM v2: SFT on synthetic + human-written rewrites.

- AdLlama (proposed): Imitation v2 + RLPF using reward model from performan

All models are based on Llama 2 Chat 7B.

6. Experiment Design (A/B Test)

- Period: Feb 16, 2024 – Apr 25, 2024 (10 weeks).

- Population: 34,849 US advertisers.

- Randomization: Advertiser-level assignment to control (Imitation v2) vs. treatr
(AdLlama).

- Metrics: Advertiser-level CTR, total clicks, impressions, number of ads, numbe
variations.

The evaluation emphasizes advertiser-level ROI rather than per-ad metrics, cor
business impact goals.

7. Main Results

- CTR improved by 6.7% relative ($p = 0.0296$) for AdLlama vs. Imitation v2.

- Absolute CTR: approximately 3.1% → 3.3%.

- Ad variations increased by 18.5%, suggesting higher adoption and satisfactio

- Number of ads did not change, implying the effect is mainly in better text, no
ads.

These are meaningful gains in a mature ad platform where large improvements
achieve.

8. Limitations

- Offline-only RL: The reward model is trained on historical data. It does not ada
online or explore new trends.

- Single objective: Optimizes CTR only, not creativity, tone adherence, or brand
constraints.

- Human selection not modeled: Advertisers choose which AI-generated variant
reward model does not include selection probability.

- Platform-level impacts: Broader effects (e.g., ad diversity, user experience) ar
analyzed.

9. Ideas for Continuation

- Online RL loop: incorporate real-time feedback from generated ads.

- Multi-objective reward: combine CTR with tone constraints, creativity, or adve
satisfaction.

- Selection-aware reward: weight outcomes by likelihood of advertiser acceptan

- Cross-domain adaptation: apply RLPF to email marketing, product description
customer support with performance metrics.

10. Why This Paper Is Important

This paper provides one of the first large-scale, real-world evaluations of RL pos
training impact on business metrics. It demonstrates that aligning LLMs with ag
performance feedback (not just human preferences) can yield tangible economi
is a critical step in understanding how LLMs move from “capable” to “impactful