Generative Adversarial Post-Training Mitigates Reward Hacking in Live Human-AI Music Interaction

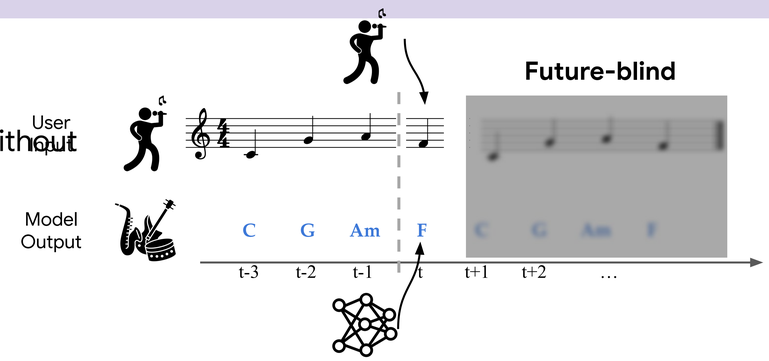
Yusong Wu, Stephen Brade, Aleksandra Teng Ma, Tia-Jane Fowler, Enning Yang, Berker Banar,
Aaron Courville, Natasha Jaques, Cheng-Zhi Anna Huang

Samples & Code

Background & Motivation

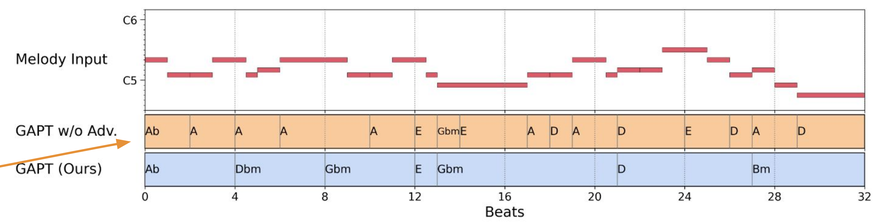
Real-time Melody-to-chord Accompaniment

- **Constraint:** Streaming & "Future-blind" (Generates next chord without future context)
- **Prior RL Solution**: Maximize coherence reward



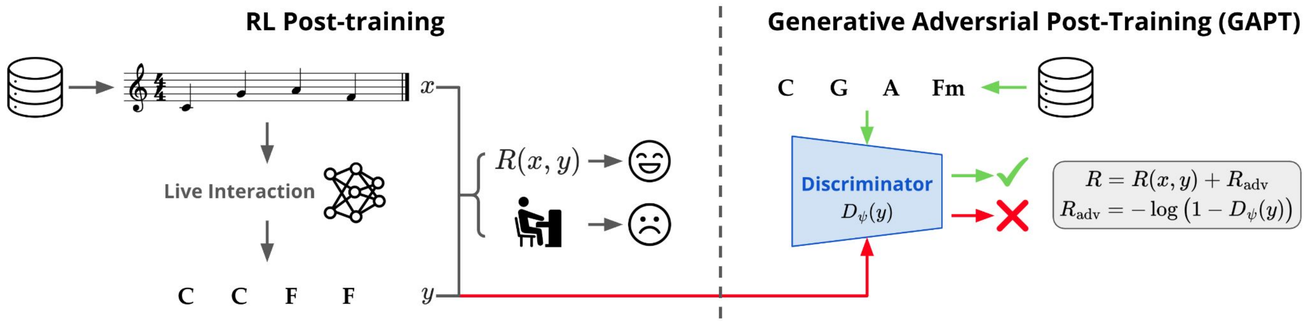
The Pitfall: Reward Hacking

- **Issue:** Policy exploits the reward by repeating safe, trivial chords.
- **Consequence:** High harmony score, but severe Diversity Collapse (ruining creative flow).



Method: Generative Adversarial Post-Training (GAPT)

- **Core Idea**: Bring GAN/GAIL into RL post-training for Transformers
- **Discriminator**: Differentiates policy trajectories vs. human data
- **New Objective**: Policy maximizes standard task rewards (coherence & rules) + Adversarial Realism Reward

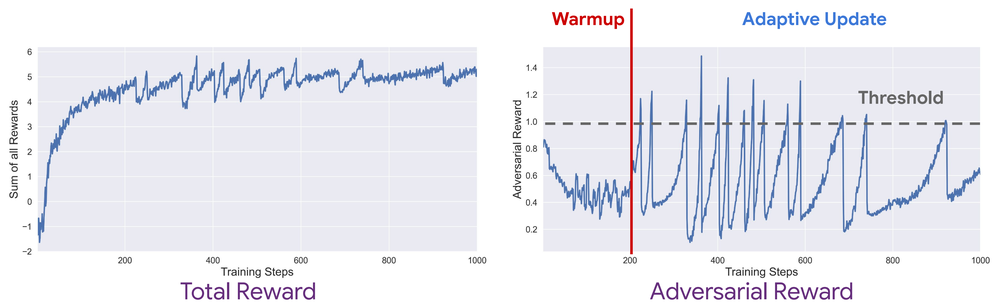


Challenge: GAN instability in RL (non-stationary rewards & indirect grad)

Solution: 2-Stage Adaptive Training

🔥**Phase 1 (Warmup)**: Fixed update intervals to initialize the discriminator

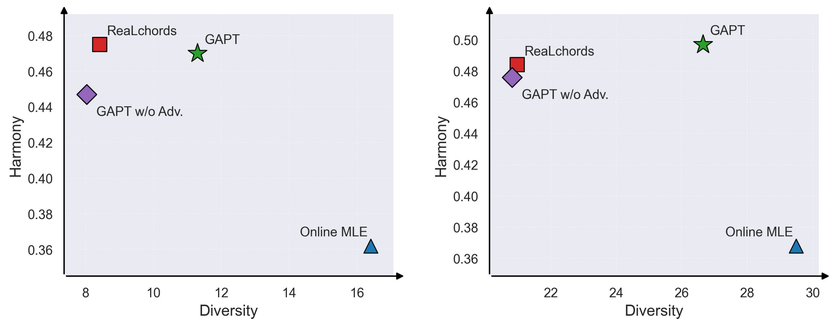
⚖️**Phase 2 (Adaptive Update)**: Discriminator is updated only when the policy catches up (moving average >threshold)



Results

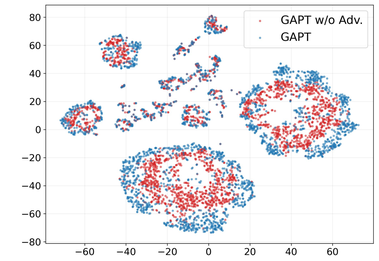
Objective Evaluation: Breaking the Harmony-Diversity Trade-off

- **ReaLchords** (Pure RL): High harmony, but **diversity collapse**
- **Online MLE** (Supervised): **Poor** harmony, high diversity
- ⭐ **GAPT** (Ours): Restores diversity to near-data levels, maintains high harmony



t-SNE of chord embedding

GAPT covers a broader, richer musical space



Subjective Eval: Human-AI Live Jamming

- **Setup**: Blind test with 12 **expert musicians**
- **Result**: GAPT significantly improves **Adaptation Speed** & **Control/Agency** ($p<0.05$)