RL-Based Finetuning (Reinforcement Learning Finetuning)

Reinforcement Learning (RL) based finetuning refers to training a model (e.g., policy network, trading agent, or LLM) using rewards obtained from interacting with an environment rather than (or in addition to) supervised loss.

In Stockify, this can mean:

- Fine-tuning a trading policy to maximize cumulative returns or Sharpe ratio.
- Finetuning a sentiment classification model to align better with downstream trading performance.

Key Elements

- **Agent (Policy Network)**: The trading model that chooses portfolio weights.
- **Environment**: Market simulation or real-time trading API.
- **Reward Signal**: Portfolio returns, risk-adjusted return, drawdown, etc.
- **States**: Market conditions, sentiment scores, news embeddings, etc.
- **Actions**: Asset allocation decisions.

Human-in-the-Loop (HITL)

HITL methods combine human feedback with automated learning to improve model performance, safety, and interpretability. In financial applications, HITL is essential due to the high-risk, high-stakes nature of decisions.

Let's explore common HITL-RL paradigms and how they apply to Stockify:

• RLHF – Reinforcement Learning with Human Feedback

- Use human experts (analysts or financial advisors) to rank outputs of the trading agent or sentiment model.
- Train a **reward model** from these rankings.
- Use **PPO (Proximal Policy Optimization)** to finetune the policy to maximize the learned reward.

Application:

- Rank predicted sentiments or portfolio allocations.
- Train a reward model to mimic expert preference.
- Finetune trading policy to align with expert-judged good outcomes.

- **RLAIF – Reinforcement Learning with AI Feedback**

Like RLHF but uses another model instead of a human to provide feedback.

In Stockify:

- Use a strong LLM (like GPT-4) to act as an evaluator of trading decisions or sentiment classification, scoring outputs based on explainability, relevance, and alignment with financial goals.

Advantages:

- Scalable.
- Less expensive than RLHF.
- Faster iteration during back testing.

Example:
Use a finance-tuned GPT model to evaluate if a news-based sentiment aligns with expected price movement. Use its judgment as a reward function for RL-based fine-tuning of the sentiment classifier.

- **RLSF – Reinforcement Learning from Simulated Feedback**

Like RLAIF but the feedback comes from a simulated environment (e.g., trading simulator or market model).

In Stockify:

- Build a market simulator.
- Observe how certain predictions (e.g., sentiment-driven trades) perform.
- Use simulated outcomes as feedback.

Application:

- Evaluate how a sentiment classifier affects downstream portfolio return.
- Reward models that contribute to profitable decisions.

- **CriticGPT – Critique-based Finetuning**

OpenAI's CriticGPT critiques model outputs and flags issues for training correction.

In Stockify:

- Build or use a “CriticLLM” trained to evaluate:
    - Poor sentiment outputs.
    - Risky trades.
    - Misaligned portfolio allocations.
- Use it to provide critiques of agent behavior, not just scalar rewards.
- This can help produce explanations and corrections during fine-tuning.

Example:
CriticGPT says: “This sentiment classification ignores contradictory news,” or “This trade increases volatility risk.”

We can use these critiques to fine-tune your sentiment or trading models.

- **HRLAIF – Human-Reviewed LLM-Augmented AI Feedback**

This is a hybrid of RLHF and RLAIF, where:

- An LLM generates feedback.
- A human reviews/approves/refines that feedback.
- Used to ensure high-quality, domain-aligned feedback in sensitive domains like finance.

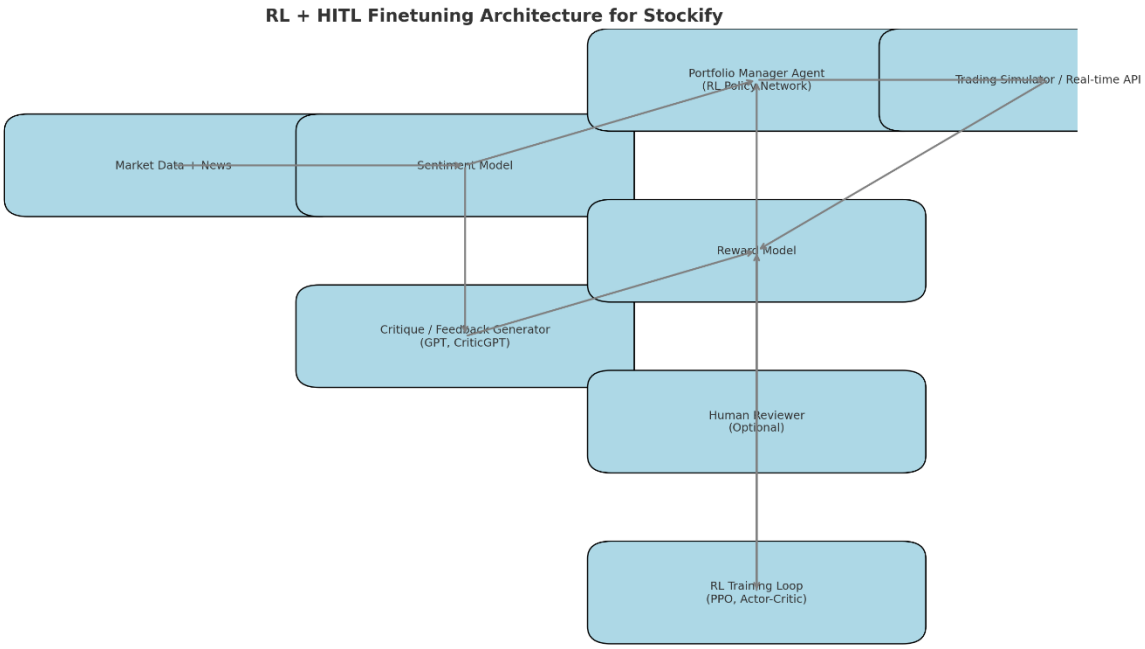
In Stockify:

- Use GPT-like models to evaluate financial summaries or trades.
- Have a domain expert (trader or analyst) verify or edit the feedback.
- Train the reward model on these verified scores.

Example:
GPT rates a portfolio 7/10 due to overexposure to tech. A human adjusts it to 5/10 due to missing macro risks. This corrected score is used for reward model training.

Below is the architecture for stockify using RL and HITL Finetuning

-



- **Market Data + News**: Feeds raw financial inputs.
- **Sentiment Model**: Classifies and scores sentiment from news/headlines.
- **Portfolio Manager (RL Agent)**: Allocates assets based on sentiment + market state.
- **Trading Simulator/API**: Tests actions in a simulated or live environment.
- **Critique Generator (e.g., GPT, CriticGPT)**: Evaluates sentiment and strategy outputs.
- **Reward Model**: Learns to assign scores based on outcomes or critiques.
- **Human Reviewer**: Optional step to refine LLM feedback.
- **RL Training Loop**: Uses feedback (e.g., PPO) to update trading strategy.