# CTDE Simple Spread Training Report

This report summarizes the results of the Centralized Training with Decentralized Execution (CTDE) implementation on the PettingZoo MPE 'simple\_spread' environment with communication.

## 1. Implementation Overview

The training used a CTDE approach where each agent has its own policy network for decentralized execution, while a centralized critic has access to the joint observations of all agents for training. Communication between agents was implemented using a discrete message channel of size K\_vocab=4.

Key Components Implemented:  
- SimpleSpreadCTDE environment wrapper to handle actions and messages.  
- CommPolicy for agent policy and message generation.  
- CentralCritic for centralized value estimation.  
- PPO-style training loop for both actor and critic.  
- Logging of performance metrics in CSV.  
- Analysis script to produce plots of learning curves and performance improvement.

## 2. Training Configuration

Main configuration parameters:  
- NUM\_EPISODES: 1000  
- MAX\_STEPS: 50  
- K\_VOCAB: 4  
- LR\_ACTOR: 3e-4  
- LR\_CRITIC: 3e-4  
- GAMMA: 0.95  
- GAE\_LAMBDA: 0.95  
- CLIP\_RATIO: 0.2  
- UPDATE\_EPOCHS: 4  
- ENTROPY\_COEF: 0.01  
- MSG\_ENTROPY\_COEF: 0.01  
- VAL\_COEF: 0.5

## 3. Results

The results are presented using two plots:  
1. learning\_curve.png — Learning curve showing the mean return per episode.  
2. first\_last\_bar.png — Comparison of mean return in the first 100 vs last 100 episodes.

Observations:  
- The learning curve shows high variance and noisy returns, with no clear upward trend.  
- The bar plot comparison shows minimal improvement from first to last episodes, indicating the policy did not converge to significantly better performance.

## 4. Suggestions for Improvement

To improve results, consider:  
- Increasing NUM\_EPISODES (e.g., 5000+) to allow more training.  
- Lowering LR\_ACTOR and LR\_CRITIC slightly (e.g., 1e-4) to stabilize learning.  
- Increasing UPDATE\_EPOCHS to improve policy optimization.  
- Experimenting with larger K\_VOCAB for richer communication.  
- Tuning entropy coefficients to balance exploration vs exploitation.  
- Trying reward shaping or curriculum learning to guide agents early in training.