Phase 3 Documentation: Continual Reinforcement Learning with Subgoal-Conditioned Policies

Author: Rishav Tewari

Date: 17-01-2026

Abstract This document presents a detailed and reproducible account of Phase 3 of the project. Phase 3 integrates the preference-aligned language model developed in Phase 2 into a continual reinforcement learning setting. The objective of this phase is to evaluate whether subgoal-conditioned policies, guided by a frozen and aligned planner, can mitigate catastrophic forgetting and improve sample efficiency when learning sequential tasks. The phase emphasizes controlled experimentation, strict separation of learning components, and careful evaluation across a fixed task curriculum.

Objectives of Phase 3 Phase 3 is designed to address the following research questions:

- Can a frozen, preference-aligned language model serve as an effective high-level planner in a continual reinforcement learning setting.
- Does subgoal conditioning reduce catastrophic forgetting compared to flat reinforcement learning baselines.
- How does rehearsal-based replay affect long-term retention across tasks.

Concretely, Phase 3 has the following objectives:

- Integrate the Phase 2 planner into a hierarchical RL agent without updating the planner parameters.
- Train a low-level PPO policy sequentially across a fixed curriculum of MiniGrid tasks.
- Compare subgoal-conditioned agents against flat PPO and scripted subgoal baselines.
- Quantify forgetting, transfer, and adaptation speed under controlled conditions.

System overview Phase 3 follows a hierarchical reinforcement learning architecture composed of two strictly separated components.

High-level planner:

- Implemented using the Phi-2 language model.
- Loaded with LoRA adapters trained in Phase 2.
- Operates in inference-only mode throughout Phase 3.

- Produces symbolic subgoals conditioned on textual environment state.

Low-level policy:

- Implemented as a PPO agent.
- Receives subgoals as part of its observation space.
- Learns to execute subgoals through interaction with the environment.
- Is the only component updated during training.

This separation ensures that improvements in performance can be attributed to policy learning rather than planner adaptation.

Task curriculum Phase 3 uses a fixed, ordered curriculum of MiniGrid environments selected to progressively increase planning complexity.

The curriculum is as follows:

- MiniGrid-DoorKey-6x6-v0
- MiniGrid-DoorKey-8x8-v0
- MiniGrid-Unlock-v0
- MiniGrid-KeyCorridorS3R1-v0
- MiniGrid-ObstructedMaze-2Dlh-v0

Tasks are trained strictly sequentially. The agent is evaluated on all previously seen tasks after completing training on each new task.

Planner integration, the planner is instantiated at the start of Phase 3 and remains fixed. Planner integration follows these principles:

- The planner receives a textual description of the current environment state.
- It outputs a symbolic subgoal drawn from the canonical grammar defined in Phase 1.
- Invalid or unparseable outputs are handled through deterministic fallback logic.
- Planner inference does not depend on gradients or learning signals.

The planner is treated as an external decision module rather than a trainable component.

Policy learning The low-level policy is trained using Proximal Policy Optimization. Policy learning is configured as follows:

- One PPO agent is trained across all tasks.

- Training proceeds sequentially without resetting policy parameters between tasks.
- Subgoals are embedded and concatenated with environment observations.
- PPO hyperparameters are held constant across tasks and baselines.

This setup isolates the effect of task sequencing and subgoal conditioning.

Rehearsal-based replay to mitigate catastrophic forgetting, Phase 3 introduces rehearsal-based replay at the policy level:

- Transitions from previously learned tasks are stored in replay buffers.
- During training on task k, PPO minibatches mix current-task data with a fixed fraction of data from tasks 1 to k-1.
- The planner remains frozen and does not participate in replay.

Rehearsal is applied only to the policy learning process.

Baselines Phase 3 includes the following baselines.

Flat PPO

- Standard PPO agent without a planner.
- Observes only the environment state.
- Serves as a non-hierarchical baseline.

Scripted subgoals

- Uses a deterministic, heuristic-based planner.
- Outputs symbolic subgoals using the same interface as the LLM planner.
- Provides an upper-bound heuristic comparison without learning.

These baselines are evaluated under identical training and evaluation protocols.

Evaluation protocol Evaluation is performed after training on each task:

- Agents are evaluated on all tasks seen so far.
- Performance is measured using fixed seeds.
- No learning occurs during evaluation.

Metrics The following metrics are reported:

- Episode success rate per task.
- Average return.

- Forgetting measured as performance drop on earlier tasks.
- Episodes to reach a fixed success threshold.

Results are aggregated across multiple random seeds.

Failure modes and constraints Observed constraints during Phase 3 include:

- Memory limitations restricting the use of PPO with large language models.
- Sensitivity to subgoal parsing errors.
- Trade-offs between replay buffer size and training stability.

These constraints are documented to inform future extensions.

Phase 3 outcome Phase 3 demonstrates that subgoal-conditioned policies guided by a frozen, preference-aligned planner can improve continual learning performance. The results motivate further analysis through ablation studies and scaling experiments in subsequent phases.