Phase 5 — Technical Documentation

LLM-Driven Rehearsal, Subgoal Reuse Analysis, and Cost-Aware Gating

Author: Rishav Tewari
Date: 20-01-2026

1. Overview

Phase 5 introduces the first *mechanism-level* contributions of the project. Whereas earlier phases establish feasibility (Phase 1), alignment (Phase 2), continual learning behavior (Phase 3), and component necessity via ablations (Phase 4), Phase 5 formalizes *why* the system works by introducing explicit rehearsal, reuse, and cost-control mechanisms.

This phase operationalizes the hypothesis that **language-mediated abstractions can serve as durable, reusable memory in continual reinforcement learning**, reducing catastrophic forgetting while improving adaptation speed.

2. Objectives

The primary objectives of Phase 5 are:

- Introduce a principled rehearsal mechanism based on abstract subgoals rather than raw trajectories
- Quantify transfer across tasks using a formal Subgoal Reuse Ratio
- Reduce catastrophic forgetting observed in Phase 3 baselines
- Control LLM interaction cost via a learned gating mechanism

Phase 5 is designed to produce *novel, publishable insights* rather than incremental engineering improvements.

3. Conceptual Motivation

Traditional experience replay stores low-level transitions that are tightly coupled to specific environments. Such replay often fails under distribution shift and contributes to interference in continual learning.

Phase 5 replaces low-level replay with **LLM-generated subgoal rehearsal**, where high-level intent is preserved while irrelevant stochastic detail is discarded. This abstraction enables reuse across tasks, environments, and noise regimes.

4. System Components

Phase 5 consists of five interacting subsystems:

1. **Aligned Phi-2 Subgoal Generator**
   Generates canonical, executable subgoals conditioned on environment state.

2. **Deterministic Subgoal Parser and Validator**
   Ensures syntactic and semantic executability of generated subgoals.

3. **Synthetic Rehearsal Repository**
   Stores validated subgoal traces indexed by task and seed.

1. **Continual PPO Agent**
   Executes subgoals using intrinsic reward shaping.

2. **Cost-Aware Gating Policy (Optional)**
   Controls when to query the LLM.

Each component is independently ablatable, supporting rigorous scientific evaluation.

5. Synthetic Subgoal Rehearsal

5.1 Prompt Design

The LLM is prompted using structured templates enforcing a canonical subgoal vocabulary (e.g., GOTO, PICK, OPEN, DELIVER). Prompts include textual state summaries and explicit execution constraints to minimize hallucination.

5.2 Generation Strategy

- Temperature sweep: 0.0, 0.5, 0.8
- Nucleus or top-k sampling variants
- Multiple traces generated per task seed

5.3 Validation and Filtering

Generated subgoals undergo:

- Deterministic parsing into canonical tokens
- Lightweight environment simulation
- Length and timeout filtering

Only validated traces are retained for rehearsal.

6. Replay Integration

6.1 Interleaved Replay

Synthetic subgoal traces are sampled during online PPO training and injected as intrinsic reward signals. Replay batches are weighted by a tunable replay coefficient.

6.2 Pretraining Replay

The agent performs offline gradient updates on rehearsal traces prior to learning a new task, biasing behavior toward reusable skills.

7. Subgoal Reuse Analysis

7.1 Subgoal Reuse Ratio

The Subgoal Reuse Ratio (SRR) measures the fraction of subgoals executed during new-task learning that appear in rehearsal traces from previous tasks.

7.2 Interpretability

SRR enables causal analysis by linking reuse to adaptation speed and forgetting reduction, transforming qualitative intuition into quantitative evidence.

8. Gating Policy

The gating policy determines when the LLM should be queried based on state embeddings, recent reward trends, and task identity.

Training approaches include:

- Supervised learning from Phase 4 logs
- Reinforcement learning with explicit query cost penalties

The gating policy enables efficient deployment without sacrificing performance.

9. Evaluation Protocol

All Phase 5 experiments enforce strict controls:

- Identical task curricula
- Fixed compute budgets
- Shared PPO hyperparameters
- Multiple random seeds

10. Metrics

Primary Metrics

- Subgoal Reuse Ratio

- Forgetting Score
- Episodes-to-Threshold
- Query Cost Efficiency

Secondary Metrics

- Subgoal execution accuracy
- Replay acceptance rate
- Wall-clock efficiency

11. Statistical Analysis

Results are reported as mean and standard deviation across seeds. Paired statistical tests and effect sizes are used to establish significance.

12. Failure Modes and Mitigations

- Hallucinated subgoals mitigated via constrained vocabularies and validation
- Negative transfer mitigated via replay weighting
- Excessive LLM cost mitigated via gating

13. Deliverables

Phase 5 produces:

- Runnable rehearsal and training scripts
- Reproducible configuration files
- Rehearsal datasets and LoRA adapters
- Paper-ready figures and tables

14. Role in the Full Research Program

Phase 5 serves as the *mechanistic core* of the project. It enables Phase 6 robustness evaluation, Phase 7 interpretability, and Phase 8 publication and release.

15. Summary

Phase 5 transforms LLM-guided hierarchical control from a promising idea into a defensible continual learning mechanism, providing the empirical and conceptual foundation for the remainder of the research program.