Phase 4: Ablation Studies

Author: Rishav Tewari

Dated: 17-01-2026

Overview and Scientific Motivation

Phase 4 is dedicated to systematic ablation studies designed to isolate and quantify the contribution of each architectural and algorithmic component in the continual reinforcement learning framework. The objective is to move beyond aggregate performance gains and establish causal evidence for why and when Phi-2 guided hierarchical control improves adaptation speed and mitigates catastrophic forgetting.

This phase is critical for scientific rigor and paper readiness. Well executed ablations convert an engineering system into a defensible research contribution by demonstrating necessity, sufficiency, and interaction effects among components.

Core Research Questions

The ablation experiments are structured to answer the following questions:

1. Which components are strictly necessary for continual learning gains?
2. Does alignment of the LLM matter more than its raw capacity?
3. How much of the improvement comes from hierarchy versus reward shaping?
4. What is the tradeoff between parameter efficiency and performance stability?
5. How sensitive is the system to structured versus free form language outputs?

Each ablation removes or alters exactly one factor while holding all others constant, ensuring interpretability of results.

Experimental Control Protocol

To maintain experimental validity, the following controls are enforced across all ablations:

- Identical MiniGrid task curricula and environment seeds
- Fixed compute budget measured in environment steps
- Same PPO hyperparameters unless explicitly ablated
- Same random seeds for policy initialization
- Consistent evaluation checkpoints

Any deviation from these controls is explicitly documented and justified.

Mandatory Ablation Dimensions

1. LLM Training State

Objective: Measure the effect of language model alignment on subgoal quality.

Variants:

- Fixed Phi-2 (no fine tuning)
- DPO aligned Phi-2
- RLHF aligned Phi-2

Hypothesis: Alignment improves subgoal consistency and reduces compounding errors during long horizons.

2. Alignment Method

Objective: Compare preference based alignment strategies.

Variants:

- Direct Preference Optimization
- PPO based RLHF with learned reward model

Hypothesis: DPO provides better stability under limited data, while RLHF may yield marginal gains at higher computational cost.

3. Intrinsic Subgoal Reward

Objective: Evaluate the role of hierarchical reward shaping.

Variants:

- PPO with intrinsic reward on subgoal completion
- PPO without intrinsic reward (extrinsic only)

Hypothesis: Intrinsic rewards accelerate early learning but may reduce robustness if over weighted.

4. Parameter Adaptation Strategy

Objective: Understand parameter efficiency versus plasticity.

Variants:

- LoRA adapters only
- Full model fine tuning

Hypothesis: LoRA preserves prior task knowledge and reduces forgetting relative to full fine tuning.

5. Subgoal Representation

Objective: Test the necessity of structured language.

Variants:

- Constrained canonical subgoal vocabulary
- Free form natural language subgoals

Hypothesis: Canonical subgoals reduce execution ambiguity and improve sample efficiency.

Metrics and Evaluation

Each ablation is evaluated along three orthogonal axes:

Adaptation Metrics

- Episodes to reach predefined success threshold
- Learning curve area under the curve

Continual Learning Metrics

- Per task success rate after sequential training
- Forgetting score defined as maximum performance drop relative to peak

Subgoal Quality Metrics

- Subgoal execution accuracy
- Average steps per subgoal
- Subgoal timeout frequency

All metrics are reported as mean and standard deviation across multiple seeds.

Statistical Analysis

- Paired statistical tests are used to compare ablation variants
- Effect sizes are reported alongside p values
- Confidence intervals are included for all primary metrics

This ensures claims are supported by statistically meaningful evidence rather than anecdotal trends.

Expected Outcomes and Interpretation

Phase 4 is expected to demonstrate that:

- LLM alignment is a first order contributor to continual learning gains
- Hierarchical reward shaping is necessary but not sufficient alone
- Parameter efficient adaptation via LoRA significantly reduces forgetting

- Structured subgoal interfaces outperform unconstrained language

These findings directly motivate the novel contributions explored in subsequent phases.

Deliverables

- Complete ablation result tables
- Learning curves for each ablation dimension
- Consolidated summary plots for paper inclusion
- Clear textual interpretation of each result

All artifacts are logged with experiment IDs and made reproducible.