Phase 7 — Technical Documentation

Analysis, Interpretability, and Failure Diagnostics

Author: Rishav Tewari
Date: 20-01-2026

---

1. Overview

Phase 7 focuses on **analysis, interpretability, and diagnostic understanding** of the continual learning system developed in Phases 1–6. While earlier phases establish mechanisms (LLM-guided subgoals, rehearsal, gating) and validate their robustness and transfer properties, Phase 7 explains *why* the system behaves as observed, *where* it fails, and *how* different components contribute causally to performance.

This phase transforms raw experimental results into **scientifically interpretable evidence**, which is essential for top-tier publication and internal research validation.

2. Objectives

The primary objectives of Phase 7 are:

- Interpret subgoal-level behavior across tasks, noise regimes, and domains
- Systematically categorize and quantify failure modes
- Diagnose the behavior and effectiveness of the gating policy
- Attribute robustness and generalization gains to specific mechanisms
- Produce paper-ready visualizations and qualitative evidence

Phase 7 performs **no new training**; all analysis is conducted offline using artifacts from previous phases.

3. Inputs from Previous Phases

Phase 7 explicitly consumes artifacts generated in earlier phases:

- Phase 5: synthetic subgoal traces, replay logs, reuse statistics, gating metadata
- Phase 6: robustness experiment logs, transfer rollouts, noise-conditioned evaluations
- Policy checkpoints and execution trajectories

All analysis scripts operate deterministically on these fixed inputs.

4. Analysis Structure

Phase 7 is organized along four orthogonal analysis axes, each addressing a distinct scientific question.

5. Subgoal Semantics and Reuse Analysis

5.1 Subgoal Taxonomy

- Categorize subgoals by action type (navigation, manipulation, interaction)
- Analyze object references and spatial arguments
- Measure frequency and temporal position within episodes

5.2 Reuse Decomposition

Subgoal reuse is decomposed into:

- **Exact reuse**: identical canonical subgoal tokens
- **Functional reuse**: same semantic intent with different parameters
- **Spurious reuse**: reused syntax without positive contribution

5.3 Performance Correlation

- Correlate reuse statistics with adaptation speed
- Correlate reuse statistics with forgetting and robustness metrics

6. Failure Mode Analysis

6.1 Failure Taxonomy

Failures are classified into mutually exclusive categories:

- LLM hallucination (invalid or misleading subgoals)
- Execution mismatch (valid subgoal, failed execution)
- Policy drift (policy ignores correct subgoal)
- Environment mismatch (subgoal valid in source but invalid in target domain)

6.2 Noise-Conditioned Failures

- Analyze failure distributions as a function of observation noise
- Analyze failure distributions as a function of transition noise

6.3 Degradation Dynamics

- Distinguish catastrophic collapse from graceful degradation
- Measure recovery attempts and partial successes

7. Gating Policy Diagnostics

7.1 Query Timing Analysis

- Analyze when LLM queries occur relative to decision boundaries
- Identify over-querying and under-querying regimes

7.2 Counterfactual Evaluation

- Compare gated behavior against simulated always-query and never-query baselines
- Identify states where gating prevents harmful or unnecessary queries

7.3 Cost–Stability Tradeoff

- Plot robustness versus query budget
- Identify diminishing returns in query usage

8. Representation and Causal Attribution

8.1 Subgoal Embedding Analysis

- Embed subgoals using token-level or semantic representations
- Visualize clusters across tasks, noise levels, and domains

8.2 Causal Replay Ablations

- Remove specific subgoal categories from rehearsal data
- Measure resulting performance degradation

These analyses attribute gains to concrete subgoal families.

9. Phase-7-Specific Metrics

Phase 7 introduces new diagnostic metrics:

- Failure type frequency
- Graceful degradation index
- Subgoal semantic entropy
- Query harm rate
- Recovery success rate

Metrics are reported per task, per noise level, and per seed.

10. Visualization Requirements

Mandatory Phase 7 figures include:

- Subgoal reuse heatmaps across tasks

- Failure taxonomy distributions
- Robustness degradation curves with failure overlays
- Gating decision timelines
- Qualitative episode trace diagrams
- Subgoal embedding visualizations (e.g., t-SNE or UMAP)

These visualizations are critical for paper acceptance.

11. Statistical Analysis

- Report mean and standard deviation across seeds
- Use paired statistical tests where applicable
- Report effect sizes and confidence intervals

No post-hoc metric invention is permitted.

12. Reproducibility and Auditability

- Every analyzed episode is linked to experiment ID and random seed
- Analysis scripts are deterministic and versioned
- Qualitative examples are sampled automatically, not manually selected

13. Deliverables

Phase 7 produces:

- Offline analysis scripts
- Structured diagnostic datasets (CSV/JSON)
- Publication-quality figures (PDF/SVG)
- Qualitative failure appendix
- Phase 7 README and analysis report

14. Role in the Full Research Program

Phase 7 provides the **interpretability and causal evidence** required to support claims made in earlier phases. It enables Phase 8, which focuses on paper writing, artifact packaging, and open-source release.

15. Summary

Phase 7 converts performance numbers into understanding. By exposing internal structure, failure dynamics, and causal mechanisms, it elevates the project from an empirical system to a defensible scientific contribution.