
AGENTS SAFETY, EVALUATION AND ALIGNMENT FEBRUARY 12, 2026

LHAW: CONTROLLABLE UNDERSPECIFICATION FOR LONG-HORIZON TASKS

George Pu, Michael S. Lee, Udari Madhushani Sehwal, David Lee, Bryan Zhu, Yash Maurya, Mohit Raghavendra, Yuan Xue, Sam Denton

LHAW introduces a dataset-agnostic pipeline for generating and validating underspecified long-horizon tasks to evaluate how frontier models detect missing information, decide when to clarify, and recover performance under ambiguity.

Long-horizon workflow agents that operate effectively over extended periods are essential for truly autonomous systems. Their reliable execution critically depends on the ability to reason through ambiguous situations in which clarification seeking is necessary to ensure correct task execution. However, progress is limited by the lack of scalable, task-agnostic frameworks for systematically curating and measuring the impact of ambiguity across custom workflows. We address this gap by introducing LHAW (Long-Horizon Augmented Workflows), a modular, dataset-agnostic synthetic pipeline that transforms any well-specified task into controllable underspecified variants by systematically removing information across four dimensions – Goals, Constraints, Inputs, and Context – at configurable severity levels. Unlike approaches that rely on LLM predictions of ambiguity, LHAW validates variants through empirical agent trials, classifying them as outcome-critical, divergent, or benign based on observed terminal state divergence. We release 285 task variants from TheAgentCompany, SWE-Bench Pro and MCP-Atlas according to our taxonomy alongside formal analysis measuring how current agents detect, reason about, and resolve underspecification across ambiguous settings. LHAW provides the first systematic framework for cost-sensitive evaluation of agent clarification behavior in long-horizon settings, enabling development of reliable autonomous systems.

LINKS

<https://arxiv.org/pdf/2602.10525v1>