AI Assistants Overassist

Verona Teo
Stanford University, Stanford, California, United States

Raghav Jain
University of California, San Diego, San Diego, California, United States

Tobias Gerstenberg
Stanford University, Stanford, California, United States

Max Kleiman-Weiner
University of Washington, Seattle, Washington, United States

Abstract

Large language models (LLMs) are increasingly being used as tutors and thought partners, yet how they navigate intervention decisions during problem-solving remains poorly understood. While AI guidance can scaffold learning, its benefits depend on how such systems help—intervening too early or too frequently may hinder learning and cognitive engagement. Here, we introduce Int-Bench, a simulation-based benchmark for evaluating LLM interventions. Int-Bench simulates a "student" solving a problem while a "teacher" monitors the student's reasoning and decides whether, when, and how to intervene. Across three domains—code debugging, mathematics, and brain teasers—we evaluate LLM teachers on intervention frequency and timing, and their impact on immediate task success and generalization to new problems. Compared to humans, LLMs intervene more frequently and earlier, and tend to provide complete solutions rather than targeted hints. These findings suggest that current LLM assistants often optimize for short-term success rather than preserving the reasoning processes needed for learning.

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*