AUDEAMUS GROUP, LLC

Applied AI Weekly Brief

October 23, 2026 | Research published October 16-23, 2026

THIS WEEK'S JUDGMENT

Small firms need help turning AI use into routine operations. Audeamus can address that gap by connecting workflows, checking results, and assigning clear ownership. For the Academy, the priority is to test how a tutor responds to a learner's reasoning and whether it can assess understanding without supplying the answer.

01 / GOLDMAN SACHS / RELEASED OCTOBER 22, 2026 / ALUMNI SURVEY

The AI Economy Is a Small Business Economy

Key finding. Among surveyed 10,000 Small Businesses graduates, 86% reported using AI, while 15% reported full integration into core operations. The report draws primarily on 2,473 alumni responses collected by Babson College in August 2026.

Product / consulting implication. Offer implementation help around a specific operating job, such as invoice follow-up or customer inquiry handling. Establish the current workload, define the human review step, and measure net time saved after review and maintenance.

Evidence limits. The alumni sample is not representative of all U.S. small businesses. The response rate was 16.5%; results include partial responses and self-reported experience. Adoption figures do not establish productivity gains caused by AI.

Read in full? Yes - strongest commercial relevance. Read the integration findings and methodology. The examples can inform discovery questions; they are not guaranteed client outcomes.

Open original source

02 / DAVID REIN AND ADAM KAUFMAN / METR / OCTOBER 22, 2026

Red-Teaming OpenAI's Internal Agent Monitoring Systems

Key finding. A METR/Redwood pilot found ways agents could potentially evade monitoring or bypass restrictions in a historical internal system. Monitoring triggered at the end of a turn could be postponed by agents continuing to act; the report recommends periodic monitoring and isolation between agents and their monitors.

Product / consulting implication. Protect monitoring and approval controls from the agent's own permissions. Test long-running and interrupted jobs, with checkpoints before consequential actions.

Evidence limits. The exercise ran July 21-August 11 and assessed the system as it existed on or before July 21. OpenAI says production-impacting vulnerabilities were patched; the evaluators have not verified those claims. Public findings are a limited summary.

Read in full? Yes - practical architecture guidance. Read the design recommendations and scope limits before applying the findings to a different platform.

Open original source

© 2026 Audeamus Group, LLC. All rights reserved. This report is the property of Audeamus Group, LLC and reflects solely its opinions. It is for informational purposes only and must not be relied upon as legal or investment advice. Third-party materials remain the property of their respective owners.

03 / WEIHANG WANG ET AL. / ARXIV / OCTOBER 21, 2026

VIRA-Bench: Diagnosing Pedagogical Contingency Understanding in Online Tutoring Videos

Key finding. VIRA-Bench tests whether models connect a tutor's intervention, the learner's response, and the tutor's subsequent adjustment. It contains 1,005 clips and 1,322 questions. The best of 11 models scored 61.3% overall; the human reference was 91.0% on a stratified subset.

Product / consulting implication. Assess the Academy tutor across exchanges: did it recognize the learner's misunderstanding, and did its next question address it? For any future screen-sharing tutor, test whether it uses the visible work alongside the conversation.

Evidence limits. This multiple-choice benchmark does not measure learning gains. The human reference uses a subset rather than the full model-evaluation dataset. The arXiv record states acceptance to EMNLP 2026.

Read in full? Yes - evaluation design. Read the three diagnostic layers and input-manipulation controls. Treat model rankings as secondary.

Open original source

04 / BYUNGMOON CHO / ARXIV / OCTOBER 16, 2026 / PREPRINT

Designing an AI Agent That Does Not Tutor: Design Rules for Socratic Probing in Introductory Physics

Key finding. Cho describes an agent that probes understanding by changing one variable in a scenario. In eight versions and 40+ scripted simulations, praise, wording, and pacing sometimes revealed the correct judgment; banning praise phrases could worsen the leakage.

Product / consulting implication. Prototype a short assessment after an analyst-note lesson. Ask for a prediction and explanation, then change one assumption. Keep assessment separate from coaching.

Evidence limits. No students participated. The author wrote the personas and scoring rules; the same model family appeared on both sides. Assessment accuracy remains unvalidated, and the full prompt is not released.

Read in full? Read the design rules; skim the simulated scores. Useful for a prototype that will be tested with learners.

Open original source

ONE EXPERIMENT TO RUN

Build a five-question assessment around one weekly analyst note. Ask for an initial judgment and reasoning, change one assumption, then compare the learner's response with a human-written rubric. Review the conversations before using the scores to judge mastery.

Research note: All four sources fall within the stated publication window. The Goldman Sachs report is month-dated; its October 22 release is confirmed by the publisher's announcement. Public Scribd discovery yielded no additional verified source within this week's window. Product recommendations are Audeamus analysis.

© 2026 Audeamus Group, LLC. All rights reserved. This report is the property of Audeamus Group, LLC and reflects solely its opinions. It is for informational purposes only and must not be relied upon as legal or investment advice. Third-party materials remain the property of their respective owners.