Proceedings of the International Conference on Industrial Engineering and Operations Management

Publisher: IEOM Society International, USA DOI: 10.46254/AF6.20250214
Published: April 8, 2025

Practical AI Benchmark: Evaluating Integrated Reasoning, Multimodal Analysis, and Tool Use in Industrial Contexts

Zakaria Zaza and Omar Souissi
EFFYIS-GROUP
INPT
Rabat, Morocco

Abstract

Artificial intelligence (AI) have rapidly evolved from stand-alone text-only large language models (LLMs) into multi-capable agents with integrated reasoning, tool use, and multimodal abilities (e.g., GPT-4, Grok, Google Gemini). These advanced systems promise to tackle complex real-world tasks, but evaluating their performance in practical industrial scenarios remains challenging. Existing benchmarks such as GAIA [1], BFCL [2], and MageBench [3] each focus on a subset of required skills (general question-answering, function calling, or multimodal reasoning) and do not reflect the integrated challenges of industrial applications. To address this gap, we present the PracticalAI Benchmark, a comprehensive evaluation suite for AI systems in realistic industrial environments. PracticalAI tasks demand a combination of capabilities: multi-step logical reasoning for problem-solving, tool usage via API calls to interact with simulated sensors and actuators, processing of image and video inputs, and web search integration to verify real-time technical data. Unlike traditional evaluations that emphasize response correctness, PracticalAI prioritizes execution accuracy – measuring whether an AI agent can autonomously execute the correct sequence of tool calls to accomplish each task. All tasks run in high-fidelity simulated industrial environments, ensuring safe yet realistic testing of an agent's ability to plan and act autonomously. Although developed for industrial engineering use cases (e.g., manufacturing processes), the PracticalAI framework is broadly applicable to other domains requiring autonomous agents, including IT operations, software system management, and logistics.

Keywords

Multimodal AI, Industrial automation, Tool use, Multi-step reasoning, Autonomous agents, AI benchmarking.