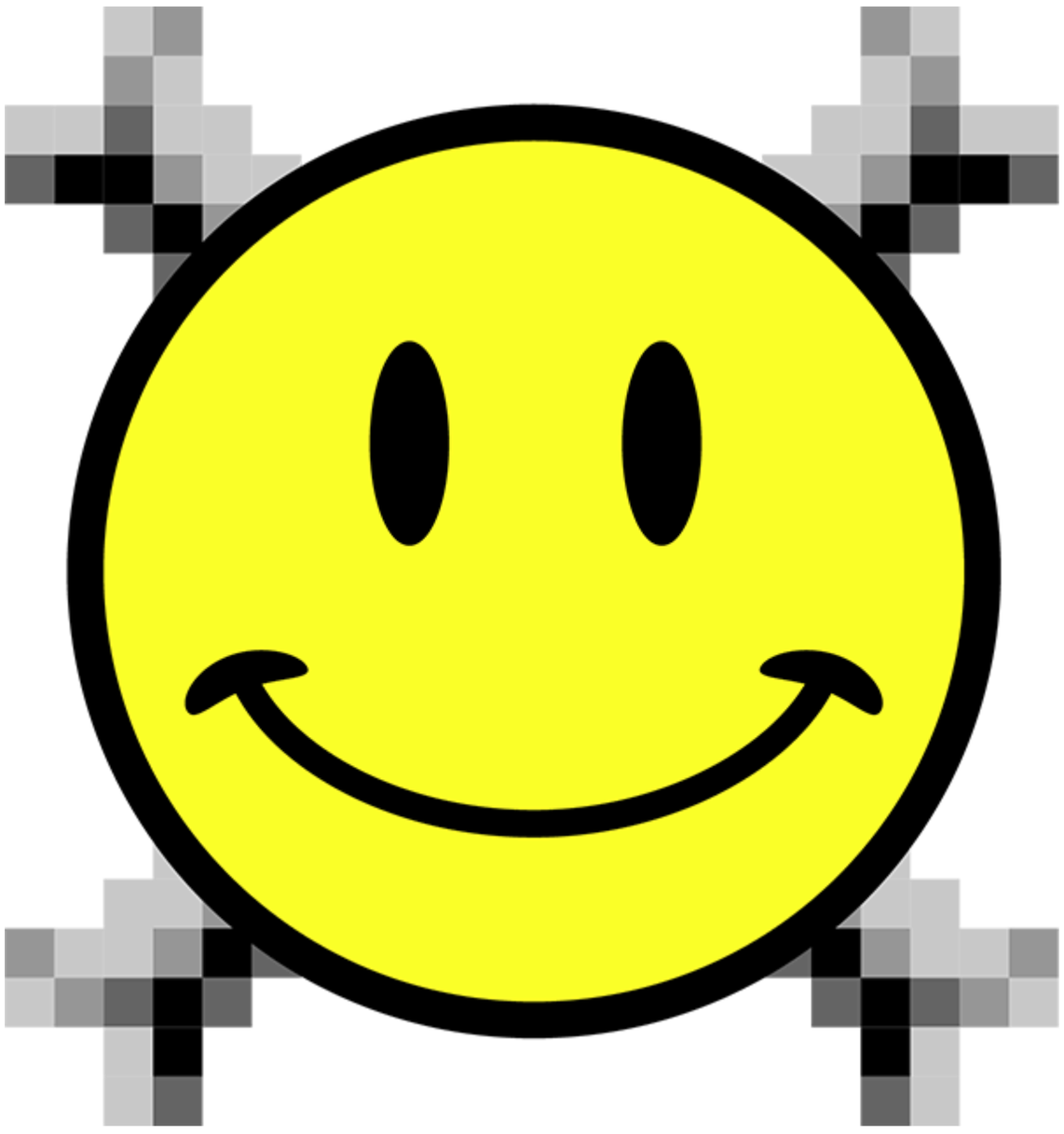

arXiv Is Hiring a DevOps Engineer

Work on one of the world's most important websites and make an impact on open science.

View Jobs

Skip to main content

Computer Science > Artificial Intelligence

arXiv:2502.15840 (cs)

[Submitted on 20 Feb 2025]

Vending-Bench: A Benchmark for Long-Term Coherence of Autonomous Agents

Axel Backlund, Lukas Petersson

View PDF HTML (experimental)

While Large Language Models (LLMs) can exhibit impressive proficiency in isolated, short-term tasks, they often fail to maintain coherent performance over longer time horizons. In this paper, we present Vending-Bench, a simulated environment designed to specifically test an LLM-based agent's ability to manage a straightforward, long-running business scenario: operating a vending machine. Agents must balance inventories, place orders, set prices, and handle daily fees - tasks that are each simple but collectively, over long horizons (>20M tokens per run) stress an LLM's capacity for sustained, coherent decision-making. Our experiments reveal high variance in performance across multiple LLMs: Claude 3.5 Sonnet and o3-mini manage the machine well in most runs and turn a profit, but all models have runs that derail, either through misinterpreting delivery schedules, forgetting orders, or descending into tangential "meltdown" loops from which they rarely recover. We find no clear correlation between failures and the point at which the model's context window becomes full, suggesting that these breakdowns do not stem from memory limits. Apart from highlighting the high variance in performance over long time horizons, Vending-Bench also tests models' ability to acquire capital, a necessity in many hypothetical dangerous AI scenarios. We hope the benchmark can help in preparing for the advent of stronger AI systems.

Subjects: **Artificial Intelligence (cs.AI)**
Cite as: **arXiv:2502.15840 [cs.AI]**
(or **arXiv:2502.15840v1 [cs.AI]** for this version)
https://doi.org/10.48550/arXiv.2502.15840

Focus to learn more
arXiv-issued DOI via DataCite

Submission history

From: Lukas Petersson [view email]
[v1] Thu, 20 Feb 2025 15:52:29 UTC (6,458 KB)