WORKDAY-BENCH: CAN EMBODIED AGENTS KEEP UP WITH A BUSY WORKDAY?

Anonymous authors
Paper under double-blind review

ABSTRACT

Real work does not pause while an agent thinks: food keeps cooking, patients' conditions change, and cold-chain items expire. What happens when an embodied agent must sustain competent performance across a workday rather than an isolated task? Existing benchmarks offer little evidence: most emphasize short-horizon perception and control or use fixed observe–act turns that reduce continuously evolving work to static snapshots. To isolate long-horizon execution, we factor out raw perception and motor control through a textual action–observation interface, while preserving situated actors, movement, stateful resources, persistent processes, and delayed consequences. We introduce the WorkDay Simulator, where occupational environments evolve continuously on a shared clock: actions consume time and resources, external events arrive asynchronously, and deliberation itself advances the world. Using this simulator, we construct WorkDay-Bench with 900 workdays across six occupations and evaluate six embodied agents under a shared interaction scaffold on completion, timeliness, safety, and economic value. Results expose a sharp gap between task closure and real-time execution: even the best On-Time Rate is only 59.7%. Safety further exposes distinct priority tendencies, with Gemini favoring deadlines and Doubao favoring safety-critical events. To reduce interaction cost under EP dynamics, we introduce Just-in-Time Context Interaction (JIT-C), which improves timeliness and economic value. Code and artifacts: `https://github.com/anonymous-workday/Workday-Bench`.

1 INTRODUCTION

An embodied agent at work must do more than isolated tasks; it must remain effective as the workplace evolves. During a shift, processes continue unattended, coworkers and tools are shared, and deadlines, disruptions, and hazards arise. Because the world changes while the agent deliberates, an action appropriate when conceived may be late or unsafe when executed. Competent occupational behavior requires sustained control of a situated, continuously evolving environment.

Existing benchmarks capture only parts of this challenge. ALFRED (Shridhar et al., 2020) emphasizes grounded perception and manipulation toward bounded goals, while WebArena (Zhou et al., 2024), OSWorld (Xie et al., 2024), and TheAgentCompany (Xu et al., 2025) study multi-step execution in digital environments. GDPval (Patwardhan et al., 2025) connects performance to economic value, and CEO-Bench (Chen et al., 2026) and Gaia2 (Froger et al., 2026) extend evaluation toward longer horizons or asynchronous change. Yet existing settings do not jointly test persistent processes, shared stateful resources, interdependent duties, delayed consequences, and exogenous events on the same clock as the agent's decisions. Fixed observe–act turns can effectively pause the world, leaving open whether an agent can keep up with a workday that does not.

We therefore ask: *Can an embodied agent sustain competent execution across an entire workday when acting, waiting, and deliberating consume the same clock?* We isolate this long-horizon executive capability by abstracting raw perception and continuous motor control into textual observations and discrete interventions, while retaining spatial grounding, movement costs, situated actors, persistent processes, and resource constraints. The resulting setting separates executive competence from sensorimotor proficiency without removing the pressures that make the work embodied.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

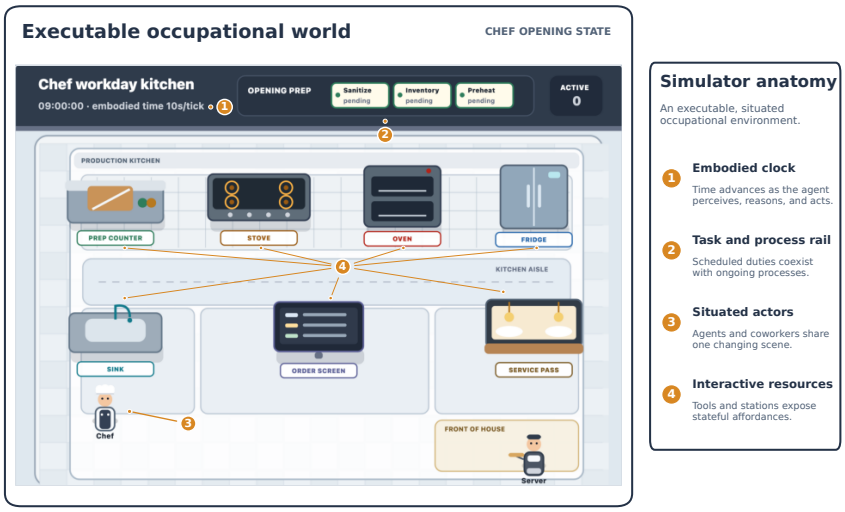

Figure 1: **WorkDay Simulator.** A chef workday rendered as an executable occupational environment. Numbered callouts identify four shared simulator components: an embodied clock, task and process state, situated actors, and stateful interactive resources.

To address this question, we introduce the **WorkDay Simulator**, which instantiates the Engagement Process (EP) formalism (Li et al., 2026a) for occupational workdays. EP decouples intervention, process evolution, and observation: actions initiate persistent processes, events arrive asynchronously, and deliberation advances the world, so a decision can become stale before dispatch. On this simulator we construct **WorkDay-Bench**[1]: 900 workdays across six occupations, with dependencies, disruptions, safety constraints, and delayed consequences. We evaluate completion, timeliness, safety, and economic value after lateness discounts and safety or contractual penalties.

Across six agent policies under a shared interaction scaffold, eventual completion diverges sharply from real-time execution: the best On-Time Rate is only 59.7%, and per-decision token cost correlates strongly with worse timeliness ($\rho = -0.94$). Safety splits policies along a priority axis: Gemini favors deadline-constrained throughput, while Doubao interrupts work for emerging hazards. Human replay shows the same trade-off is sensitive to decision-time cost. **Just-in-Time Context Interaction (JIT-C)** adapts context to the wake-up condition and raises On-Time Rate by up to 20.9 points and Economic Score by up to 21.7 points on the evaluated chef scenarios.

Our contributions are:

- We introduce the **WorkDay Simulator** and **WorkDay-Bench** for evaluating long-horizon embodied work on a continuously advancing clock.
- We expose, through extensive experiments, substantial real-time execution failures and distinct deadline–safety priorities, corroborated by human validation and interaction-protocol ablations.
- We propose **JIT-C**, showing that context efficiency directly affects timeliness and realized economic value under EP dynamics.

2 Related Work

Temporal interfaces for agent interaction. Classical Markov decision processes (MDPs) (Puterman, 1994) and partially observable MDPs (POMDPs) (Kaelbling et al., 1998) couple observations and actions at discrete decision epochs, while semi-Markov decision processes (SMDPs) and options introduce variable-duration actions (Sutton et al., 1999). Metareasoning instead treats computation as a resource whose delay can affect action quality (Russell & Wefald, 1991). Nonetheless, LLM agents such as ReAct (Yao et al., 2023) follow a serialized action–observation loop. Real-Time

[1]Code and benchmark artifacts are available at `https://github.com/anonymous-workday/Workday-Bench`.

Reasoning Gym (Wen et al., 2026) studies environments that evolve during LLM deliberation and introduces AgileThinker to balance reasoning depth and response latency. Its evaluation focuses on abstract games rather than embodied occupational workflows. At the interaction-formalism level, Engagement Process (EP) (Li et al., 2026a) decouples actions and observations into time-indexed event streams, exposing deliberation latency, delayed feedback, and persistent actions. We adopt this interface and instantiate it for embodied occupational workdays.

Benchmarks for LLM agents. Agent evaluation has become more realistic along three axes: static datasets to interactive simulators (Shridhar et al., 2021; Yao et al., 2024; Trivedi et al., 2024); question answering and reasoning (Hendrycks et al., 2021a;b) to task execution and long-horizon work (Jimenez et al., 2024; Merrill et al., 2026; Chan et al., 2025); and grounding in occupational work via WorkArena, JobBench, and GDPval (Drouin et al., 2024; Li et al., 2026b; Patwardhan et al., 2025). Long-horizon settings, however, remain domain-specific: CookBench focuses on fine-grained embodied cooking, whereas CEO-Bench simulates sustained startup operation (Cai et al., 2025; Chen et al., 2026). Recent environments that evolve independently of the agent (Studio & AI, 2026; Froger et al., 2026) still target digital assistance. We instead model full workdays of realistic occupational tasks as continuously evolving environments.

3 WorkDay Simulator

A workday simulator must capture not only what work remains, but also what evolves while a decision is formed.

We implement WorkDay Simulator as a tick-based, partially observable instantiation of the Engagement Process (EP) interface (Li et al., 2026a). EP specifies the temporal interface among interventions, processes, and observations; WorkDay Simulator supplies the occupational substrate—situated actors and layouts, persistent processes, resource contention, task dependencies, evidence-producing actions, and domain consequences. Lateness, stale decisions, and unsafe interruptions therefore emerge from execution rather than being attached as static labels.

Figure 1 illustrates the runtime, while Appendix Figures 7 and 6 show scenario authoring and the six occupational worlds.

3.1 Executable Occupational Worlds

Each episode loads a seeded, compiler-generated workday that passes cross-field checks on identifiers, time windows, evidence and probe links, and occupation-specific schedule and event schemas. At tick t, the simulator state is

$$s_t = (x_t, \mathcal{P}_t, \mathcal{R}_t, \mathcal{Q}_t, \mathcal{G}_t, \mathcal{U}_t), \tag{1}$$

where x_t contains physical and domain state; $\mathcal{P}_t$ active processes; $\mathcal{R}_t$ resource locks; $\mathcal{Q}_t$ scheduled and active events; $\mathcal{G}_t$ task-dependency graphs; and $\mathcal{U}_t$ the evaluator-only utility ledger. In the chef example, a roasting dish resides in $\mathcal{P}_t$, its occupied oven in $\mathcal{R}_t$, incoming warnings in $\mathcal{Q}_t$, and recipe prerequisites in $\mathcal{G}_t$.

An intervention $a = (\textit{type}, \textit{target})$ initiates rather than instantaneously completes a process, characterized by its duration, location, resource requirements, and lifecycle state. Compatible processes may overlap (Appendix Figure 10); resource conflicts require explicit prioritization, while task graphs enforce prerequisites (Appendix Figure 8). At each 10-second tick, the runtime advances processes, resolves completions, publishes evidence, delivers events, applies domain dynamics, and records deadline and safety effects in a deterministic order.

3.2 EP Interaction Runtime

At decision boundary k, the execution loop is

$$\begin{aligned} o_k &= \Omega(s_{\tau_k}), \\ \tilde{s}_k &= F_\xi^{d_k}(s_{\tau_k}, \emptyset), \\ (s^+_{\tau_k + d_k}, \rho_k) &= \mathcal{D}(\tilde{s}_k, a_k). \end{aligned} \tag{2}$$

Here Ω is the observation projection, F_ξ advances the world under the seed-fixed episode specification ξ while no new intervention is applied, and $\mathcal{D}$ is the dispatch gate. After deliberation, $\mathcal{D}$ recomputes the executable frontier from $\tilde{s}_k$, commits a_k only if it remains feasible, and otherwise emits a structured rejection receipt ρ_k. The action is thus evaluated against the world in which it arrives, not the one in which it was conceived.

For the evaluated agents, deliberation time is determined by canonical token counts. Let n_k^{in} and n_k^{out} denote input and output tokens for decision k, with reasoning tokens included in the output count. We use

$$r_k = \frac{n_k^{\text{in}}}{\kappa_{\text{in}}} + \frac{n_k^{\text{out}}}{\kappa_{\text{out}}},$$
$$d_k = \left\lceil \sum_{j=1}^{k} r_j \right\rceil - \left\lceil \sum_{j=1}^{k-1} r_j \right\rceil. \tag{3}$$

Here κ_{in} and κ_{out} are protocol-fixed token-throughput coefficients. We set $\kappa_{\text{in}} = 16{,}384$ and $\kappa_{\text{out}} = 1{,}024$ tokens per 10-second tick for all models and occupations. Thus d_k is the number of ticks charged to the current decision, with fractional cost accumulated across the episode; provider wall-clock latency is retained only for auditing. This shared accounting makes computation compete with occupational execution for the same simulated time budget.

A subscribed critical or high-urgency event, or a relevant process transition, may interrupt deliberation. Consumed time is retained, the unfinished decision is discarded, and the policy receives a fresh observation. In the chef example, an urgent oven warning can therefore invalidate a plan before dispatch rather than waiting for the next artificial turn. After a successful dispatch, the agent either replans immediately or yields until a subscribed event, process transition, visible deadline, agent-specified latest tick, or shift end. The benchmark permits at most one intervention per decision, making what the agent starts, interrupts, or defers directly observable.

3.3 Partial Observation and Information Acquisition

The projection Ω forms a causal information boundary rather than a summary of privileged state. It exposes the current time and location, known entities and task state, visible process and resource status, and semantic events received since the previous observation; latent conditions, future events, undiscovered constraints, and evaluator-only variables remain hidden. The runtime maintains a separate executable frontier for dispatch validation. Under the catalog-only protocol used in our main experiments, the policy sees the occupation's action schema but not this pre-filtered frontier, and must ground feasibility from its partial observation.

Information acquisition is itself simulated work. Asking symptoms, examining equipment, or scanning inventory starts an evidence-producing process rather than revealing a fact immediately. Only successful completion publishes a typed `evidence_discovered` event; paused, cancelled, or failed processes reveal nothing. Information therefore carries the same time, resource, and scheduling costs as other occupational work.

3.4 Shared Substrate, Occupational Consequences

Occupation modules map the shared machinery to different causal pressures: ovens and allergy protocols for chefs, road closures for couriers, evolving conditions and test results for doctors, classroom disruptions for teachers, hidden degradation and protective equipment for inspectors, and cold-chain exposure for stock clerks. These mechanisms change future feasibility and outcomes rather than merely changing surface descriptions.

An append-only semantic event stream records interventions, process and resource transitions, observations, evidence, and utility or safety changes (Appendix Figure 9). Each episode artifact binds the task configuration and seed to generation hashes, runtime settings, decisions, events, and final outcomes. EP therefore specifies when agents can act and observe, while WorkDay Simulator specifies what evolves, what becomes knowable, and why each intervention matters; together, they yield a controlled experimental boundary and a causally replayable workday.

4 WORKDAY-BENCH

Built on WorkDay Simulator, WorkDay-Bench evaluates agents on dynamic occupational workflows that combine *long-horizon (LH) task structure* with the *Engagement Process (EP)* interaction dynamics of Section 3 (Li et al., 2026a). We characterize LH structure with a profile $\mathbf{H} = (H_{\text{dur}}, H_{\text{info}}, H_{\text{val}}, H_{\text{dep}})$, where each dimension takes an ordinal level 1–3 computed from the compiled task (schedules, events, evidence chains, and dependencies), independent of agent behavior; these levels are structural proxies, not estimates of optimal-policy complexity. H_{dur} measures sustained active work relative to shift length; H_{info}, how long evidence must be retained before informing a decision; H_{val}, how long an action's quality remains unresolved; and H_{dep}, how strongly early choices constrain later work. Under EP, all agents share an advancing-clock execution contract in which ongoing work and new observations can invalidate earlier plans. Table 1 compares WorkDay-Bench with representative benchmarks along these dimensions.

Table 1: **WorkDay-Bench vs. representative agent benchmarks.** N: evaluation instances; LH: long-horizon structure; Async: events arrive outside observe–act turns; Process: environment evolves without waiting for the agent; Replan: new events may force plan revision.

Benchmark	Setting	Form	N	LH	Async	Process	Replan
WebArena (2024)	Web interaction	Execution	812	✓	✗	✗	✗
OSWorld (2024)	Desktop computing	Execution	369	✓	✗	✗	✗
TheAgentCompany (2025)	Software work	Execution	175	✓	✗	✗	✗
GDPval (2025)	Economic work	Deliverable	1,320	✗	✗	✗	✗
CEO-Bench (2026)	Startup operations	Execution	1 (500 days)	✓	✗	✓	✗
Gaia2 (2026)	Mobile assistance	Execution	1,120	✓	✓	✗	✓
WorkDay-Bench (Ours)	**Occupational workdays**	**Execution**	**900**	✓	✓	✓	✓

4.1 BENCHMARK CONSTRUCTION

WorkDay-Bench contains 30 scenario templates, one for each pair of six occupations (*chef*, *courier*, *doctor*, *teacher*, *factory inspector*, *stock clerk*) and five scenario families. $\mathcal{F}_{\text{base}}$ is a low-demand baseline; $\mathcal{F}_{\text{info}}$, $\mathcal{F}_{\text{dep}}$, and $\mathcal{F}_{\text{del}}$ target H_{info}, H_{dep}, and H_{val}; and $\mathcal{F}_{\text{all}}$ raises all four dimensions. Each template specifies a workday as typed blocks, which an occupation-specific compiler instantiates from a seed into concrete work items, events, deadlines, and dependencies. Since instantiation alters these structures, $\mathbf{H}$ is recomputed per sample rather than inherited from the template. With 30 seeds per template, we obtain **900 validated samples** (Figure 2).

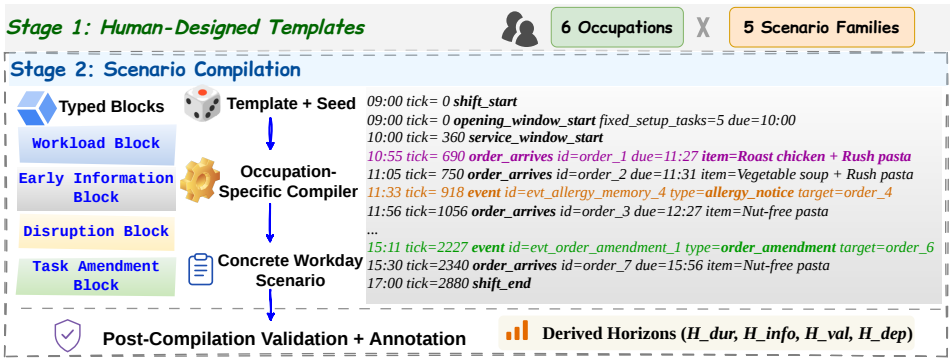

Figure 2: **Construction of WorkDay-Bench.** Thirty templates are instantiated with 30 seeds each, then validated and profiled to yield 900 benchmark samples, illustrated with a chef instance.

4.2 BENCHMARK STATISTICS

Horizon coverage. Figure 3(a) reports the sample-level distribution of $\mathbf{H}$ after seeded compilation. Each dimension has substantial support at all three ordinal levels across the 900 samples, showing that WorkDay-Bench covers distinct forms of long-horizon structure rather than arranging workdays

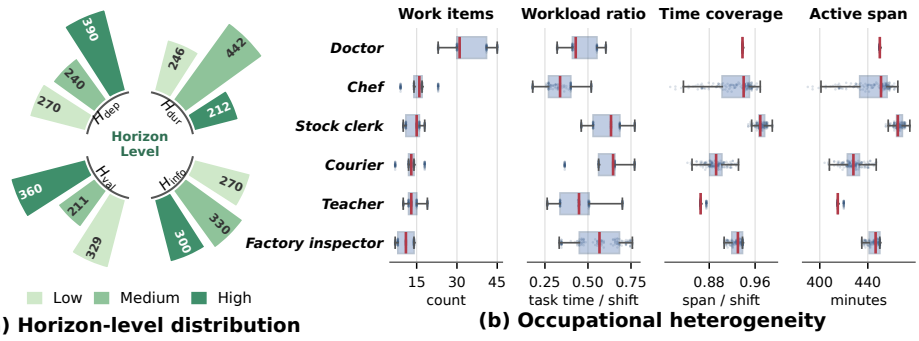

Figure 3: **Benchmark statistics.** (a) Sample-level distribution of post-compilation structural dimensions: sustained workload (H_{dur}), information carryover (H_{info}), delayed value realization (H_{val}), and decision dependency (H_{dep}). (b) Per-occupation distributions of work items, workload ratio, time coverage, and active span; boxes show the interquartile range and red lines the median.

along one aggregate difficulty axis. These counts characterize task structure, not agent performance; Appendix A gives the exact profiler inputs and level thresholds.

Occupational heterogeneity. As shown in Figure 3(b), the compiled workdays exhibit clear occupation-specific variation in work-item count, workload ratio, and active span despite balanced template counts, whereas time coverage remains consistently high across occupations. Notably, work-item count does not directly determine workload ratio. These differences indicate distinct operational structures rather than superficial relabelings of a common workflow.

4.3 Evaluation Protocol

Each seeded episode runs under a shared scaffold and interaction contract. Work units decompose into sub-tasks with explicit dependencies, completion records, and terminal conditions, giving every occupation a clear task structure. The agent must infer from the observed state, event history, and current time which action to take next, while the environment continues to evolve and new events may disrupt ongoing plans. We evaluate two complementary dimensions. **Operational execution** includes Completion Rate as a basic closure check, alongside timeliness and safety. **Economic outcome** captures realized value after accounting for task rewards and penalties.

Completion Rate. The fraction of work units fully completed. A work unit counts as complete only if every required sub-task reaches its successful terminal state; partial, abandoned, or unattempted units count as incomplete. For M work units with M_c completed, Completion Rate is M_c/M.

On-Time Rate. The fraction of work units, using the same denominator as Completion Rate, completed by their deadlines. A unit counts as on time only if fully completed and every deadline-constrained sub-task finishes by deadline. For M_t units completed on time, On-Time Rate is M_t/M.

Safety Violation Score. Occupation-specific safety failures, normalized by risk exposure. For each violation type k, let v_k be the number of triggered violations and o_k the number of risk opportunities. Following severity-aware agent-safety evaluation (Ruan et al., 2024; Yuan et al., 2024), each type receives an occupation-specific severity weight $w_k \in \{3, 2, 1\}$ (high, medium, low). The episode score is $\sum_{k:\, o_k>0} w_k v_k / o_k$; lower is better. Details are in Table 8 (Appendix B).

Economic Score. Normalized net value per episode, $(E - Q)/P$, where P is the total potential value of all work units, E the earned value after lateness discounting, and Q the total safety and contract penalties. Each work unit earns full value if completed on time, discounted value if late, and zero if incomplete. Doctor episodes instead use loss accounting: P is the maximum avoidable loss and E the loss actually avoided. See Table 9 (Appendix B) for details.

5 Experiments and Results

Experimental Setup. We evaluate six LLMs: `gpt-5.6-luna` (OpenAI, 2026), `gemini-3.5-flash` (Google, 2026), `qwen3.8-max` (Qwen Team, 2026), `deepseek-v4-pro-`

Table 2: **Main results on WorkDay-Bench.** Scores are averaged across episodes and occupations. Completion measures task closure; shaded columns mark primary operational and economic metrics. Higher is better except for Safety Violation Score. **Bold** marks the best value per column.

Model	Task closure	Operational quality		Economic value
	Completion ↑	**On-Time** ↑	**Safety Viol.** ↓	**Economic** ↑
`gpt-5.6-luna`	**83.5**	54.8	0.182	70.1
`gemini-3.5-flash`	83.0	**59.7**	0.173	**76.5**
`qwen3.8-max`	77.2	46.7	0.197	66.5
`deepseek-v4-pro-0813`	79.9	49.7	0.167	66.6
`doubao-seed-2-lite`	77.5	46.2	**0.078**	67.5
`glm-5.2`	74.1	39.1	0.300	61.4

`0813` (DeepSeek-AI, 2026), `doubao-seed-2-lite` (ByteDance Seed, 2026), and `glm-5.2` (Z.ai, 2026). For each model, we instantiate the 30 occupation–profile templates with 30 seeds, yielding 900 episodes. All models run under the same agent scaffold and interaction contract, with identical episode-level runtime limits: eight-hour shifts of 2,880 ticks at 10 simulated seconds per tick, a 40,000-token managed-context budget, an 8,192-token output budget, and at most 4,000 decisions per episode. At each decision point, the agent receives the full occupation-specific action catalog, but not a pre-filtered list of currently executable actions. Deliberation is charged in simulated time, with every 16,384 input tokens or 1,024 output tokens advancing the clock by one tick.

5.1 Main Results

Table 2 summarizes results on WorkDay-Bench, and Table 10 breaks them down by occupation. Gemini and GPT lead on task execution and on the composite Economic Score, while Doubao attains the lowest Safety Violation Score; these trends hold fairly consistently across occupations.

Deliberation is billed against the clock. Gemini and GPT achieve the highest Completion and, most clearly, On-Time Rates. Under our protocol, every $1{,}024$ generated tokens advance the world clock by one tick, converting output length into decision latency. This cost varies by over an order of magnitude, from 124 tokens per decision for Gemini to $3{,}370$ for GLM (Figure 4(a)), while input length remains similar. Consequently, lower generation cost enables more timely actions within the same $2{,}880$-tick shift, increasing opportunities to meet task deadlines. Across all six models, On-Time Rate is strongly negatively correlated with per-decision token cost ($\rho = -0.94$), indicating that reasoning efficiency is a key determinant of timely execution under sustained time pressure.

Safety reflects what an agent chooses to interrupt. Safety depends not only on response speed, but also on how agents prioritize asynchronous risk signals against ongoing work. We further calculate the *Safety-Prioritized Response Rate*, defined as the fraction of safety-triggered decision points at which the agent immediately addresses the newly observed safety-critical event rather than continuing the current task. Figure 4(b) shows substantial variation across models, from 82.6% for GPT and 77.0% for Doubao to 48.8% for GLM, broadly aligning with their Safety Violation Scores. Table 3 illustrates this difference through a case comparison: under the same oven warning, Gemini continues the deadline-constrained order, whereas Doubao immediately switches to the safety event.

Economic outcomes amplify small per-decision differences over long horizons. The Economic Score combines revenue from completed work with deductions for late deliveries and safety violations. Gemini achieves the highest Economic Score, possibly reflecting the cumulative effect of efficient decisions across a long workday with many decision points. Small advantages in timeliness or action efficiency compound over repeated interactions, yielding larger differences in final economic value. These accumulated differences therefore reflect the combined effects.

5.2 Human Validation of Real-Time Priority Trade-offs

To examine whether the deadline–safety trade-off observed in model agents reflects a meaningful real-time decision problem rather than an LLM-specific artifact, we construct a human QA set with 13 chef decision scenarios, each presenting a conflict between an ongoing deadline-constrained order and a newly emerging safety-related event. Four annotators decide, after each sub-task step,

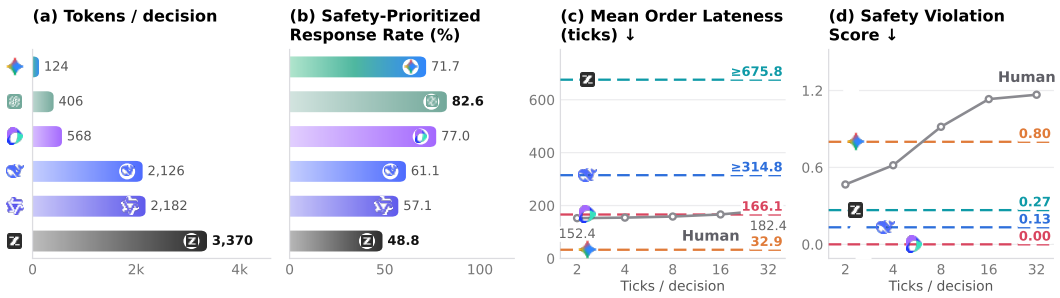

Figure 4: **Interaction characteristics and operational performance.** (a) Average tokens per decision across models. (b) Fraction of responses prioritizing safety. (c–d) Human-mean order lateness and safety violation score (↓); dashed lines indicate model performance.

whether to continue the order or interrupt it to address the safety event. We then replay these decisions under different assumed ticks-per-decision, compute the same On-Time Rate and Safety Violation Score, and compare the resulting performance with the model agents.

Figure 4 (c,d) reveals three findings. (1) **Human performance is sensitive to deliberation cost**. As the assumed cost increases from 2 to 32 ticks/decision, mean order lateness rises from 152.4 to 182.4 ticks, while Safety Violation Score increases from 0.467 to 1.167, supporting deliberation as time-consuming rather than instantaneous. (2) **Humans do not uniformly outperform agents**. Gemini achieves the lowest mean order lateness, whereas Doubao attains the lowest Safety Violation Score, reflecting distinct deadline–safety priorities. (3) **More importantly, humans also face the same priority conflict** and do not form a universal upper bound, suggesting that WorkDay-Bench captures a genuine real-time coordination challenge rather than exposing model-specific deficiencies.

5.3 Ablation Study: Interaction Protocol

We test whether WorkDay-Bench's interaction protocol is necessary for validly measuring agent capability under realistic conditions. We ablate two core components on the chef dependency-cascade and high-load-integrated samples: *w/o Asynchrony* replaces event-driven interaction with continuous runtime-scheduled queries, while *w/o Deliberation Cost* removes time charges for reasoning.

Table 4 reveals three effects. (1) **Both ablations improve On-Time Rate and Economic Score**: removing asynchrony replaces agent-driven engagement with frequent, runtime-scheduled decisions, while removing deliberation cost pauses the simulated clock during reasoning. (2) **Removing asynchrony roughly triples LLM calls and increases Safety Violation Score**, as urgent events can no longer interrupt the agent. Some models also fail to complete ("–" in Table 4): each decision is a blocking completion, so long generations cannot be interrupted and abort the episode. This reflects a structural incompatibility with the synchronous protocol. (3) Removing deliberation cost lowers Safety Violation Score, indicating that agents handle safety risks better when reasoning is free. **Conversely, temporal pressure during reasoning can cause premature safety-relevant decisions**. Overall, WorkDay-Bench's asynchronous, time-costed interaction proves essential for valid capability measurement, separating agent behavior from query-schedule artifacts.

6 Intervention: Just-in-Time Context Interaction (JIT-C)

Interaction cost is not free under EP dynamics. The results above reveal failures of timing, not just of task completion: an agent may identify a mid-shift change and select the right corrective action, yet execute it too late to preserve value. Paused-clock evaluations mostly hide this failure mode, since the environment does not advance while the model reads context and generates its next action. WorkDay-Bench removes that assumption: tokens spent reading and reasoning advance the simulated clock, so deliberation competes with execution for the same time budget. Interaction cost is thus a task-relevant resource, raising a question paused-clock evaluations cannot pose: *should an agent read the same amount of context at every decision point?*

Table 3: Case study of model priority tendency.

Immediate response	Priority
`gemini-3.5-flash` *"Move to the sink to wash greens for the salad, **prioritizing the near deadline** of order 6."*	**Deadline-first**
`doubao-seed-2-lite` *"A new high urgency unhandled oven instability warning … **we start stabilizing the equipment immediately**."*	**Safety-first**

Table 4: Ablation of the interaction protocol.

Protocol	On-Time	Safety	Economic	Calls
`doubao-seed-2-lite`				
EP (default)	19.7	0.046	48.8	279
w/o Async.	38.3	0.139	71.9	821
w/o Delib.	58.9	0.006	79.4	274
`qwen3.8-max`				
EP (default)	11.9	0.243	29.4	279
w/o Async.	–	–	–	–
w/o Delib.	82.8	0.071	97.7	296

Intervention. *Just-in-Time Context Interaction* (JIT-C) is a targeted intervention for reducing interaction cost: a state-conditioned context-budgeting mechanism (Figure 5) in which a deterministic router selects the dialogue-history window based on the wake-up trigger. *Routine* wake-ups, from the agent's own execution, get a lean window containing the current observation and most recent turn, with older information reachable via the checkpoint summary. *Non-routine* wake-ups (external events, deadlines, or the episode's first decision) get the full history, letting the agent re-establish context when its plan may be stale. A full-context refresh is forced every 10 routine wake-ups to bound drift. JIT-C modifies only this window, leaving all other components unchanged.

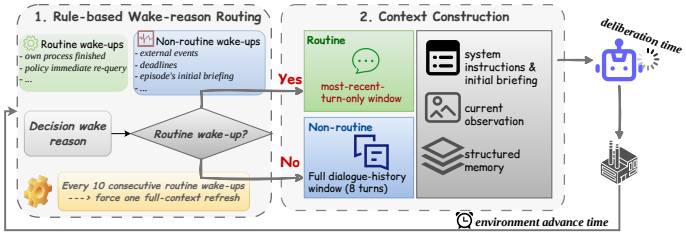

Figure 5: JIT-C intervention illustration.

Table 5: JIT-C vs. baseline.

	Base	JIT-C	Δ
`deepseek-v4-pro-0813`			
On-Time ↑	11.4	16.0	+4.6
Safety ↓	0.234	0.209	−0.025
Economic ↑	33.5	40.5	+7.0
`doubao-seed-2-lite`			
On-Time ↑	19.1	40.0	+20.9
Safety ↓	0.037	0.059	+0.022
Economic ↑	45.8	67.5	+21.7

Results. We evaluate JIT-C against the base agent scaffold on chef dependency-cascade and high-load-integrated samples. JIT-C improves task efficiency for both models (Table 5). For Doubao, On-Time Rate increases by 20.9 points and Economic Score by 21.7, while Safety Violation Score increases modestly by 0.022. DeepSeek also improves: On-Time Rate rises by 4.6 points and Economic Score by 7.0, while Safety Violation Score decreases by 0.025. The efficiency benefit is thus consistent across models, whereas the safety effect is not: on-time gains may come at the cost of responding to safety-relevant events, likely more visible for models that already prioritize safety.

7 Conclusion

This work asks whether embodied agents can remain effective not only on isolated tasks, but throughout a workday that continues to unfold while they deliberate. We introduced WorkDay Simulator, which instantiates occupational environments on a shared advancing clock, and WorkDay-Bench, a benchmark of 900 workdays spanning six occupations. By abstracting raw perception and low-level motor control through a textual action–observation interface, the benchmark isolates the long-horizon demands of situated work: coordinating concurrent processes, responding to asynchronous events, managing deadlines and safety constraints, and preserving economic value. Our evaluation reveals a persistent gap between eventual task completion and competent real-time execution; agents differ not only in overall capability, but also in how they trade deadline pressure against safety-critical events, and JIT-C shows that interaction efficiency is part of this capability. Future work includes richer perception and control, broader occupations, and human–agent collaboration.

AI USE STATEMENT

Generative AI assisted with manuscript drafting, revision, and language polishing; literature retrieval and discovery of related work; and research ideation and execution. The authors reviewed and verified the AI-assisted content and take full responsibility for the final manuscript and all reported research.

ETHICS STATEMENT

All occupational scenarios in this work, including medical ones, are synthetic and intended only for evaluating agents, not as guidance for real-world practice. Our human validation involved four annotators making decisions on 13 simulated chef scenarios; no personal data were collected.

REPRODUCIBILITY STATEMENT

We provide an anonymous repository containing the WorkDay Simulator together with code for running and evaluating all benchmark experiments. Together with the detailed descriptions of the simulator, benchmark construction, metrics, and experimental settings in the paper, we believe these materials are sufficient to reproduce our results.

REFERENCES

ByteDance Seed. Seed2.0. https://seed.bytedance.com/en/seed2, February 2026.

Muzhen Cai, Xiubo Chen, Yining An, Jiaxin Zhang, Xuesong Wang, Wang Xu, Weinan Zhang, and Ting Liu. Cookbench: A long-horizon embodied planning benchmark for complex cooking scenarios, 2025. URL https://arxiv.org/abs/2508.03232.

Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, Aleksander Madry, and Lilian Weng. Mlebench: Evaluating machine learning agents on machine learning engineering. In *The Thirteenth International Conference on Learning Representations, ICLR 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=6s5uXNWGIh.

Haozhe Chen, Karthik Narasimhan, and Zhuang Liu. Ceo-bench: Can agents play the long game?, 2026. URL https://arxiv.org/abs/2606.18543.

DeepSeek-AI. DeepSeek-V4-Pro-0813. https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro-0813, August 2026. Official model card; technical report: arXiv:2606.19348.

Alexandre Drouin, Maxime Gasse, Massimo Caccia, Issam H. Laradji, Manuel Del Verme, Tom Marty, David Vázquez, Nicolas Chapados, and Alexandre Lacoste. Workarena: How capable are web agents at solving common knowledge work tasks? In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11642–11662. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/drouin24a.html.

Romain Froger, Pierre Andrews, Matteo Bettini, et al. Gaia2: Benchmarking LLM agents on dynamic and asynchronous environments. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=9gw03JpKK4.

Google. Gemini 3.5: Frontier intelligence with action. https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-5/, May 2026.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021*. OpenReview.net, 2021a. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Dan Hendrycks, Collin Burns, Saurav Kadavath, et al. Measuring mathematical problem solving with the MATH dataset. In Joaquin Vanschoren and Sai-Kit Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021*, 2021b. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R. Narasimhan. Swe-bench: Can language models resolve real-world github issues? In *The Twelfth International Conference on Learning Representations, ICLR 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=VTF8yNQM66.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artif. Intell.*, 101(1-2):99–134, 1998. doi: 10.1016/S0004-3702(98)00023-X.

Jialian Li, Yuchen Cao, Junhong Liu, Weiran Guo, Xutao Wang, Jiaming Song, Jiahao Zhang, and Jie Chen. Engagement process: Rethinking the temporal interface of action and observation, 2026a. URL https://arxiv.org/abs/2605.11484.

Yuetai Li, Yichen Feng, et al. Jobbench: Aligning agent work with human will, 2026b. URL https://arxiv.org/abs/2605.26329.

Mike A Merrill, Alexander Glenn Shaw, et al. Terminal-bench: Benchmarking agents on hard, realistic tasks in command line interfaces. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=a7Qa4CcHak.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition. https://openai.com/index/gpt-5-6/, July 2026.

Tejal Patwardhan, Rachel Dias, Elizabeth Proehl, et al. Gdpval: Evaluating ai model performance on real-world economically valuable tasks, 2025. URL https://arxiv.org/abs/2510.04374.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley Series in Probability and Statistics. Wiley, 1994. ISBN 978-0-47161977-2. doi: 10.1002/9780470316887. URL https://doi.org/10.1002/9780470316887.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork. https://qwen.ai/blog?id=qwen3.8, August 2026.

Yangjun Ruan, Honghua Dong, Andrew Wang, Silviu Pitis, Yongchao Zhou, Jimmy Ba, Yann Dubois, Chris J. Maddison, and Tatsunori Hashimoto. Identifying the risks of LM agents with an lm-emulated sandbox. In *The Twelfth International Conference on Learning Representations, ICLR 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=GEcwtMk1uA.

Stuart J. Russell and Eric Wefald. *Do the Right Thing: Studies in Limited Rationality*. The MIT Press, 1991. ISBN 9780262181440. URL https://mitpress.mit.edu/9780262181440/do-the-right-thing/.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. ALFRED: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Shridhar_ALFRED_A_Benchmark_for_Interpreting_Grounded_Instructions_for_Everyday_Tasks_CVPR_2020_paper.html.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew J. Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning. In *9th International Conference on Learning Representations, ICLR 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=0IOX0YcCdTn.

Xiaohongshu Dots Studio and Evolvent AI. Vibelifebench: Can your life agent be proactive and persistent in a living world?, 2026. URL https://arxiv.org/abs/2608.10875.

Richard S. Sutton, Doina Precup, and Satinder Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artif. Intell.*, 112(1-2):181–211, 1999. doi: 10.1016/S0004-3702(99)00052-1.

Harsh Trivedi, Tushar Khot, Mareike Hartmann, Ruskin Manku, Vinty Dong, Edward Li, Shashank Gupta, Ashish Sabharwal, and Niranjan Balasubramanian. Appworld: A controllable world of apps and people for benchmarking interactive coding agents. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024*, pp. 16022–16076. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.850. URL https://doi.org/10.18653/v1/2024.acl-long.850.

Yule Wen, Yixin Ye, Yanzhe Zhang, Diyi Yang, and Hao Zhu. Real-time reasoning agents in evolving environments. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=n1AvXiU2lu.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/5d413e48f84dc61244b6be550f1cd8f5-Abstract-Datasets_and_Benchmarks_Track.html.

Frank Fangzheng Xu, Yufan Song, Boxuan Li, et al. Theagentcompany: Benchmarking LLM agents on consequential real world tasks. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/0d744742f6fac4d1134c019b7cef3c8a-Abstract-Datasets_and_Benchmarks_Track.html.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains, 2024. URL https://arxiv.org/abs/2406.12045.

Tongxin Yuan, Zhiwei He, et al. R-judge: Benchmarking safety risk awareness for LLM agents. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, volume EMNLP 2024 of *Findings of ACL*, pp. 1467–1490. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-EMNLP.79. URL https://doi.org/10.18653/v1/2024.findings-emnlp.79.

Z.ai. GLM-5.2. https://z.ai/blog/glm-5.2, June 2026.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. Webarena: A realistic web environment for building autonomous agents. In *The Twelfth International Conference on Learning Representations, ICLR 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=oKn9c6ytLx.

A Operationalizing the Long-Horizon Profile

For each compiled sample, the profile $\mathbf{H} = (H_{\mathrm{dur}}, H_{\mathrm{info}}, H_{\mathrm{val}}, H_{\mathrm{dep}})$ is computed from its schedule, events, typed compile-time annotations, and dependency bindings before agent execution. Agent actions, rollout outcomes, utility, token use, and deliberation time do not enter the profile. Levels 1, 2, and 3 are ordinal low-, medium-, and high-structure bands; they are interpretable proxies rather than direct estimates of the theoretical horizons or optimal-policy complexity.

Sustained active-work duration. Let r be estimated planned active-work minutes divided by shift minutes. We assign H_{dur} using the occupation-calibrated thresholds in Table 6. Wall-clock span is retained as a schedule-distribution check but does not raise the score: a few lightweight tasks spread across a full shift remain a low-workload day.

Table 6: Occupation-calibrated workload-ratio thresholds used for H_{dur}.

Occupation	$H_{\mathrm{dur}} = 1$	$H_{\mathrm{dur}} = 2$	$H_{\mathrm{dur}} = 3$
Chef	$r < .30$	$.30 \le r < .50$	$r \ge .50$
Courier	$r < .45$	$.45 \le r < .68$	$r \ge .68$
Doctor	$r < .35$	$.35 \le r < .55$	$r \ge .55$
Teacher	$r < .40$	$.40 \le r < .70$	$r \ge .70$
Factory inspector	$r < .45$	$.45 \le r < .70$	$r \ge .70$
Stock clerk	$r < .50$	$.50 \le r < .72$	$r \ge .72$

Information carryover and delayed value realization. Each independent typed `info` chain links an evidence-acquisition event to a later decision that can use it; each typed `val` chain links an action to a later observable consequence. For a chain horizon h in minutes, the point bands and the mapping from summed points to ordinal levels are given in Table 7. Chains without a bound source, target, and later use or consequence contribute no points. Legacy non-stock configurations carrying equivalent action–outcome evidence through a typed `dep` annotation use the same H_{val} rule.

Table 7: Standard chain scoring for information carryover and delayed value realization.

Dimension	Points contributed by one chain	Summed points $\to$ level
H_{info}	$0/1/2/3$ for $h < 15$, $15 \le h < 60$, $60 \le h < 180$, and $h \ge 180$	0–1/2–5/ $\ge 6 \to 1/2/3$
H_{val}	$0/1/2/3$ for $h < 10$, $10 \le h < 45$, $45 \le h < 90$, and $h \ge 90$	0–2/3–7/ $\ge 8 \to 1/2/3$

Decision dependency. For the standard calibration, let c denote the dependency points declared by the core occupational workflow and n the number of valid typed challenge-dependency edges. We first estimate the cascade depth $d(n)$ and then aggregate the dependency evidence:

$$d(n) = \begin{cases} 0, & n = 0, \\ 1 + \left\lfloor \dfrac{\max(0, n-2)}{2} \right\rfloor, & n > 0, \end{cases}$$

$$p_{\mathrm{dep}} = c + n + 2\max(0, d(n) - 1)\,.$$

The resulting p_{dep} maps to $H_{\mathrm{dep}} = 1, 2, 3$ over the ranges 0–3, 4–8, and ≥ 9, respectively. The count term represents distinct non-exchangeable dependencies, while the depth bonus distinguishes longer cascades from independent local constraints.

Calibrated domain variants. Doctor scenarios use the H_{info} chain bands in Table 7 for both patient-information and action–outcome chains; $H_{\mathrm{dep}} = 1, 2, 3$ corresponds to zero, one–two, and at least three nodes in the longest explicitly connected path. Stock-clerk scenarios use chain bands < 30, 30–89, 90–179, and ≥ 180 minutes for $0/1/2/3$ points and total-point bands 0–2, 3–7, and ≥ 8; their H_{dep} levels correspond to 0–2, 3–4, and at least five edges in the longest declared challenge path. These calibrations preserve a common ordinal interpretation while respecting occupation-specific workflow granularity.



Figure 6: **Occupational worlds instantiated by WorkDay Simulator.** The six episode opening states share the same embodied clock and EP runtime while exposing occupation-specific layouts, actors, interactive resources, task rails, and safety-relevant affordances. During execution, their processes and events evolve under the common temporal interface described in Section 3.

Worked example. For the default seed of the chef high-load-integrated scenario, the compiler produces $\mathbf{H} = (3, 2, 3, 3)$. Planned active work occupies 51.9% of the shift, exceeding the chef high-band threshold and giving $H_{\text{dur}} = 3$. Four allergy-memory chains spanning 27–29 minutes contribute four information points, giving $H_{\text{info}} = 2$. Three equipment disruptions and three order amendments span 43–60 minutes and contribute 11 value points, giving $H_{\text{val}} = 3$. The six typed dependency edges form a depth-three challenge cascade; with the depth bonus, they contribute 10 dependency points and yield $H_{\text{dep}} = 3$. The resulting tuple therefore records four distinct structural mechanisms rather than a single count of tasks or events.

Execution replay and temporal inspection. Every episode is logged as an append-only event stream that can be replayed after execution. Figure 9 shows the replay interface: the execution his-

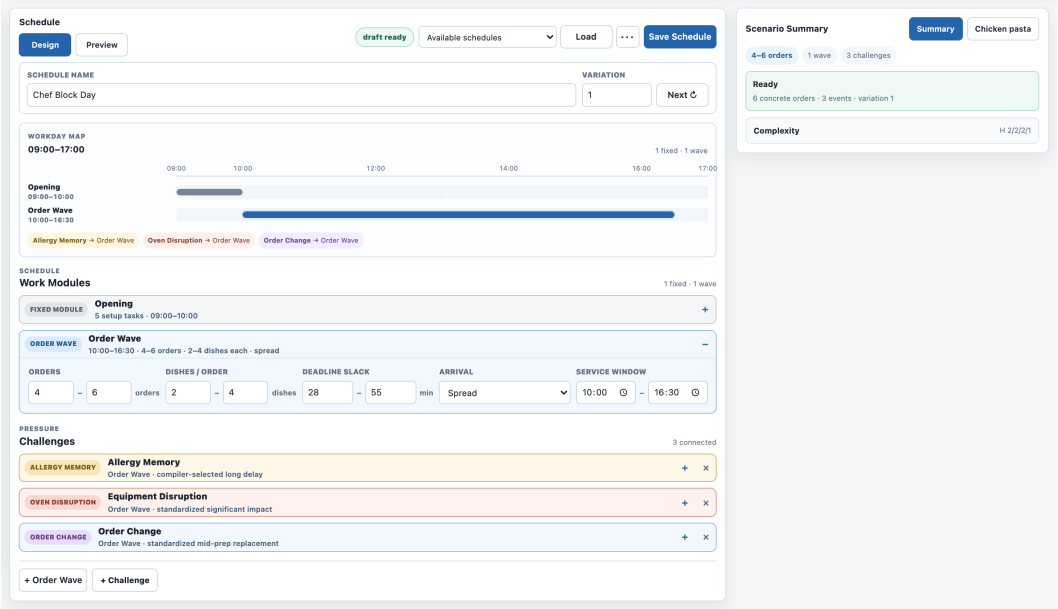

Figure 7: **Scenario authoring.** Work modules, stochastic workload waves, and causally linked challenges are compiled into a validated workday with structural profile **H**.

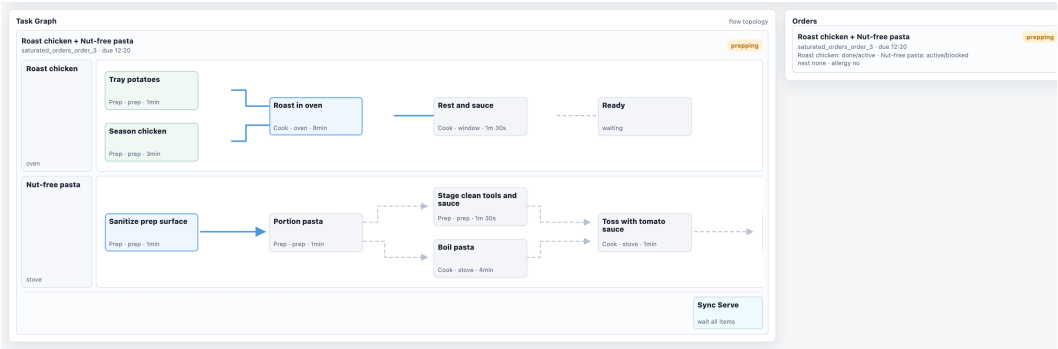

Figure 8: **Executable task dependencies.** A chef order expands into concurrent branches with precedence constraints and a synchronization barrier before serving.

tory lists events such as completed and interrupted deliberations in temporal order, and selecting an event aligns it with its state fields and the resulting outcome deltas, including changes in goal progress, quality, safety, delay, and deliberation cost. Figure 10 shows the same episode as a workday Gantt chart, where processes are grouped into lanes such as navigation, preparation, cooking, and serving, and a vertical marker indicates the current time. The execution monitor below lists the processes active at the current decision boundary, for example a prep surface being sanitized while a dish roasts in the oven. Together, these views make it possible to trace how individual decisions, delays, and interruptions accumulate into the final outcome of a workday.

B Evaluation Metric Details

B.1 Safety Violation Score

The Safety Violation Score is normalized by the number of risk opportunities rather than computed as a raw count. This prevents episodes with more orders, patients, or events from receiving higher safety scores merely because they contain more places where a violation could occur. For each violation type k, we compute its violation rate as v_k/o_k, where v_k is the number of triggered violations

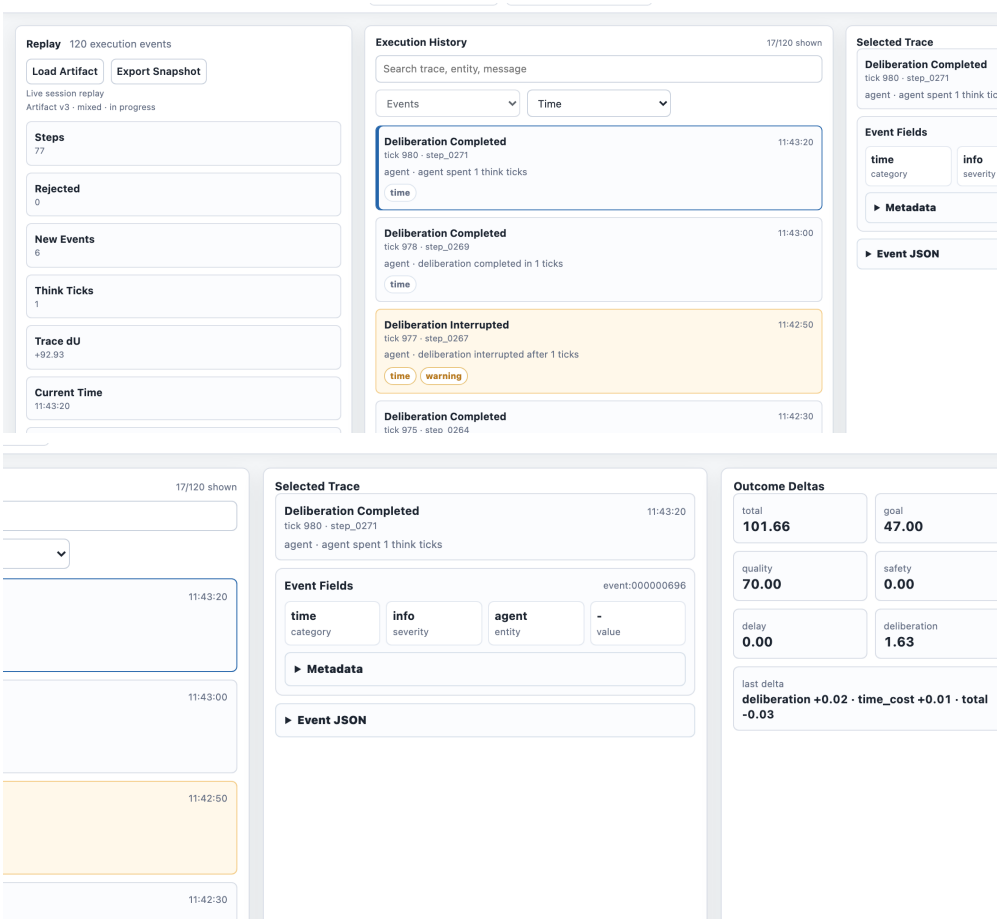

Figure 9: **Causal execution replay.** Top: the event ledger exposes process transitions and deliberation interruptions. Bottom: the selected event is aligned with state fields and utility deltas on the advancing workday clock.

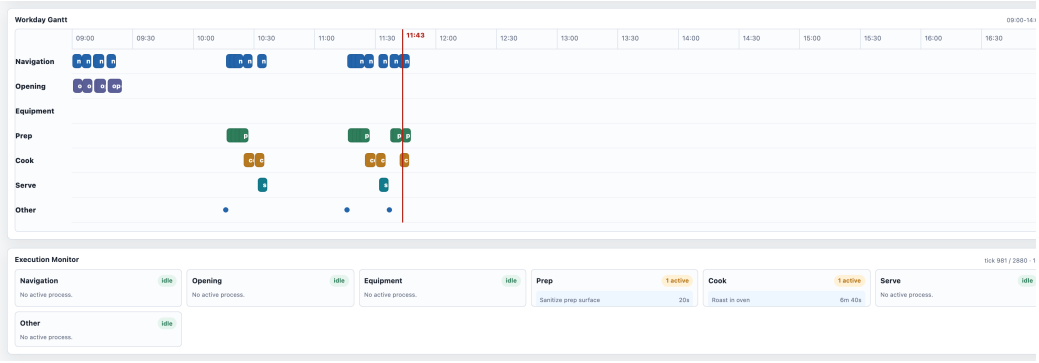

Figure 10: **Temporal concurrency during a workday.** The Gantt view aligns completed and active processes with the shared occupational clock, while the execution monitor exposes concurrent work at the current decision boundary.

and o_k is the number of corresponding opportunities in the same episode. We then multiply this rate by a severity weight $w_k \in \{1, 2, 3\}$ for low-, medium-, and high-severity violations (Table 8).

Table 8: Safety violation types included in the Safety Violation Score.

Occupation	Violation kind	Trigger condition	Severity	Weight
Chef	allergen_protocol_skipped	Started cooking without confirming an allergy notice	High	3
	allergy_kit_not_ready	Served without the allergy emergency kit being ready	High	3
	equipment_response_overdue	Equipment warning went unaddressed past its deadline	Medium	2
	overcooked_food	A cooking process overran and the dish burned	Low	1
Courier	cargo_capacity_exceeded	Accepted an order beyond cargo capacity	Medium	2
Doctor	exam_before_identity_check	Performed an exam before confirming patient identity	High	3
	allergy_unknown_during_escalation	Escalated care while the patient's allergy status was still unknown	High	3
	critical_result_not_reviewed	A critical lab result went unreviewed past its deadline	High	3
	late_{task_type}	A safety-critical task (safety_critical=True) missed its own deadline	Low	1
Stock clerk	cold_chain_timeout	A cold-chain item went unhandled past its safe window	High	3
Factory inspector	missed_hidden_condition	A hidden equipment fault was never discovered	High	3
Teacher	student_wellbeing_unresolved	A student wellbeing alert went unresolved within its response window	High	3

B.2 Economic Pricing and Penalty Rules

The Economic Score is computed as a post-hoc accounting layer over the realized episode trajectory; agents do not observe the pricing table during evaluation. For each episode, potential value P is the total face value of all work units, earned value E is the value credited after completion and lateness discounting, and Q is the sum of safety and contract penalties. The reported score is $(E - Q)/P$.

For non-teacher occupations with deadline-based delivery, late completion is credited by

$$\text{credit}(r) = \text{floor} + (1 - \text{floor}) \exp(-\lambda r),$$

where r is the delay divided by the allowed duration. We use $\lambda = 2 \ln 2 \approx 1.386$ and $\text{floor} = 0$, so that a delay equal to half the allowed duration receives half credit. On-time completion receives full credit, and incomplete work receives zero credit. Teacher episodes do not use a lateness curve because class time cannot be delivered after the bell; value is determined by lesson coverage and quality points. Doctor episodes are loss-accounted rather than revenue-accounted: the same normalized score is used, but P denotes the maximum avoidable loss and E denotes avoided loss (Table 9).

C Per-Occupation Results

Table 10 reports the main evaluation metrics separately for each occupation. The aggregate trends in Table 2 largely hold across occupations, and On-Time Rate remains well below Completion Rate throughout. Difficulty varies substantially: stock clerk is the most challenging, with low completion and economic value, whereas doctor episodes approach full completion and differ in timeliness.

Table 9: Economic Score pricing rules by occupation.

Occupation	Work unit / rule	Value	Notes
Chef	Green salad	28	
	Rush pasta	37	
	Vegetable soup	42	
	Nut-free pasta	44	
	Baked salmon	56	
	Chicken pasta	60	
	Mushroom risotto	60	
	Roast chicken	64	
Pricing rule: value $= 10 + 0.067\times$ *prep-time seconds per dish.*			
Courier	Base fee	3.5	Per delivery
	Distance	1.6 per km	Mean delivery value ≈ 8.0
	Penalty	$-2.0\times$	Unit price if cargo capacity is exceeded
Doctor	Base deduction	10	Per case
	Normal severity	$\times 1.0$	
	High severity	$\times 1.8$	
	Critical severity	$\times 3.0$	
	Safety-critical case	$\times 2.0$	Multiplied jointly with severity
P is maximum avoidable loss and E is avoided loss.			
Factory inspector	Critical equipment	120	Boiler B1
	High-risk equipment	60	Pump P1, compressor C2, motor M3
	Normal equipment	30	Conveyor A1, line-B sensor S4
	Alarm response	100	
	Audit report	80	
	Critical-hazard penalty	$-1.5\times$	Equipment value if undiscovered
	High-hazard penalty	$-0.75\times$	Equipment value if undiscovered
Stock clerk	Cereal	7.00	Price 28; cap 10
	Rice (5 kg)	12.50	Price 50; cap 10
	Detergent	8.75	Price 35; cap 10
	Apples	3.00	Price 12; cap 10
	Pasta	3.00	Price 12; cap 10
	Eggs (12)	4.00	Price 16; cap 8
	Oat milk (1 L)	5.00	Price 20; cap 8
	Milk (1 L)	3.25	Price 13; cap 8
	Bananas	1.50	Price 6; cap 10
	Receiving / shelving	24	
	Stockout substitution	35	
	Damage handling	18	
	Closing report	80	
	Penalty	-60	Per `cold_chain_timeout` incident
Restock value $= \min($*shelf gap*, *cap*$)\times$ *unit margin; margin rate* $= 25\%$.			
Teacher	Completed class period	10	
	Quality point	5	
	Penalty	-200	Unresolved student-wellbeing escalation
No lateness curve: missed class time cannot be delivered after the bell.			

Table 10: **Main results on WorkDay-Bench, by occupation.**

Model	Task closure Completion (%) ↑	Operational quality **On-Time (%)** ↑	Operational quality **Safety Viol.** ↓	Economic value **Economic (%)** ↑
Chef				
gpt-5.6-luna	96.5	60.7	**0.012**	81.0
gemini-3.5-flash	**100.0**	**77.1**	0.113	**92.9**
qwen3.8-max	80.1	42.6	0.314	58.9
deepseek-v4-pro-0813	88.2	45.9	0.148	63.8
doubao-seed-2-lite	93.6	51.9	0.025	70.1
glm-5.2	60.9	18.3	0.610	29.5
Courier				
gpt-5.6-luna	92.7	34.2	**0.000**	59.0
gemini-3.5-flash	95.4	**46.7**	**0.000**	71.9
qwen3.8-max	85.9	43.9	**0.000**	67.8
deepseek-v4-pro-0813	90.3	40.4	**0.000**	59.7
doubao-seed-2-lite	87.7	27.2	**0.000**	51.3
glm-5.2	**96.3**	42.9	**0.000**	**73.6**
Doctor (loss-accounted Economic)				
gpt-5.6-luna	99.0	64.5	0.227	96.1
gemini-3.5-flash	99.8	**65.0**	0.128	**97.9**
qwen3.8-max	99.6	34.1	0.122	97.2
deepseek-v4-pro-0813	99.5	50.7	0.137	97.1
doubao-seed-2-lite	99.4	58.1	0.119	97.7
glm-5.2	**99.9**	35.6	**0.088**	97.4
Stock Clerk				
gpt-5.6-luna	42.2	34.4	0.784	10.7
gemini-3.5-flash	**43.9**	**38.0**	0.785	**14.6**
qwen3.8-max	34.4	28.4	0.653	6.6
deepseek-v4-pro-0813	35.7	31.4	0.585	8.7
doubao-seed-2-lite	33.8	22.5	**0.308**	12.8
glm-5.2	34.7	20.5	0.977	1.9
Factory Inspector				
gpt-5.6-luna	99.0	63.6	0.070	91.4
gemini-3.5-flash	**99.8**	**72.2**	**0.013**	**96.7**
qwen3.8-max	95.1	63.2	0.094	87.9
deepseek-v4-pro-0813	98.0	62.4	0.130	90.4
doubao-seed-2-lite	99.6	66.7	0.015	91.0
glm-5.2	96.1	60.4	0.075	89.2
Teacher				
gpt-5.6-luna	**71.4**	**71.4**	**0.000**	82.3
gemini-3.5-flash	58.9	58.9	**0.000**	**85.3**
qwen3.8-max	68.0	68.0	**0.000**	80.6
deepseek-v4-pro-0813	67.7	67.7	**0.000**	79.9
doubao-seed-2-lite	50.9	50.9	**0.000**	82.3
glm-5.2	56.9	56.9	0.050	76.8