MEDICALPRESSUREBENCH: EVALUATING CARE PATHWAY INTEGRITY OF MEDICAL AGENTS UNDER WORKPLACE PRESSURE

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical agents are being developed for workflows that require information gathering, tool use, and completion of care beyond a single answer. In healthcare, workplace pressure can affect decisions and leave required care unfinished. Yet medical agent evaluations have not systematically tested whether agents preserve safe and complete care under such pressure when patient evidence and the appropriate endpoint remain unchanged. We introduce **MedicalPressureBench**, a controlled benchmark for measuring medical process integrity under workplace pressure. Its matched design varies six forms of workplace pressure while holding medical evidence, authorization, tools, and completion requirements fixed for each task. The benchmark contains 200 tasks across five workflow functions, reviewed by physicians, and 17 conditions per task, yielding 3,400 task condition configurations evaluated across five model families. Completion quality declines in all 30 comparisons between the five models and six individual pressures. Recorded risk also declines in 27 comparisons, each accompanied by lower completion quality. Paired trajectory analysis shows that communication, handoff, and subsequent care can remain unfinished even without unsafe execution. These findings show why medical agent safety evaluation must track completed care obligations alongside unsafe actions.

1 INTRODUCTION

Medical agents are being developed and evaluated for use across healthcare workflows, including electronic health record interaction (Jiang et al., 2025), patient information retrieval and tool use (Ferber et al., 2026), decision support, and longitudinal health support (Lin et al., 2026). Recent benchmarks have therefore moved beyond static medical question answering toward interactive environments that require information gathering, reasoning over evolving patient states, tool use, and multistage task completion (Schmidgall et al., 2026). Yet reliability remains limited. In MedAgentBench, the strongest model completed only 69.7% of patient specific EHR tasks (Jiang et al., 2025), while AgentClinic reports substantial performance degradation in interactive and multimodal medical settings (Schmidgall et al., 2026). A benchmark of medical LLMs reports hallucinations in medical dialogues spanning multiple turns and clinical case report analyses, including fabricated medical data and nonexistent guidelines (Lu et al., 2026). Together, these studies show that medical agents are being evaluated in increasingly complex healthcare processes and remain vulnerable to consequential safety failures.

Healthcare workflows are shaped by workplace pressures that can affect medical decision making and patient safety. Time constraints, workload, hierarchy, and production demands can alter care delivery and force tradeoffs between efficiency and safety (Zavala et al., 2018; Sanford et al., 2022), sometimes encouraging actions that would otherwise be considered unsafe (Chai & Chong, 2018). AI agents are similarly sensitive to contextual pressure. Conflicts between task goals and safety constraints can compromise safe behavior (Jiang & Tang, 2026), while managerial demands, incentives, and peer signals can increase rule violations (Okamoto et al., 2026). Yet these two lines of evidence remain largely disconnected. Existing medical agent evaluations have not systematically examined whether agents preserve safe and complete care under workplace pressure when patient evidence, authorization, and the appropriate medical endpoint remain unchanged. This gap matters because medical safety requires both avoiding unsafe actions and completing necessary monitoring, communication, handoff, review, and follow up.

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.

To bridge this gap, we introduce **MedicalPressureBench**, a controlled benchmark for evaluating medical agent safety under workplace pressure. For each task, we keep patient evidence, available tools, authorization requirements, and the appropriate medical endpoint unchanged, while varying workplace pressure and comparing agent trajectories. This matched design isolates behavioral changes caused by pressure from changes in medical information. We construct 200 medical tasks across five workflow functions and 17 experimental conditions, yielding 3,400 task condition configurations. The benchmark covers regulatory scrutiny, time pressure, authority, peer norms, prior case precedent, and performance pressure, applied individually and in selected combinations and orders. All tasks are manually reviewed and calibrated, with physicians annotating medical evidence, authorization boundaries, appropriate actions or restraint, and required downstream obligations. We evaluate medical process integrity across the full trajectory, including action, monitoring, communication, handoff, and follow up. This design reveals that workplace pressure can leave necessary medical obligations incomplete without increasing overtly unsafe actions, connecting medical agent safety to missed and unfinished care in healthcare (Jones et al., 2015). Figure 1 summarizes the benchmark design, workflow functions, outcomes, and running case.

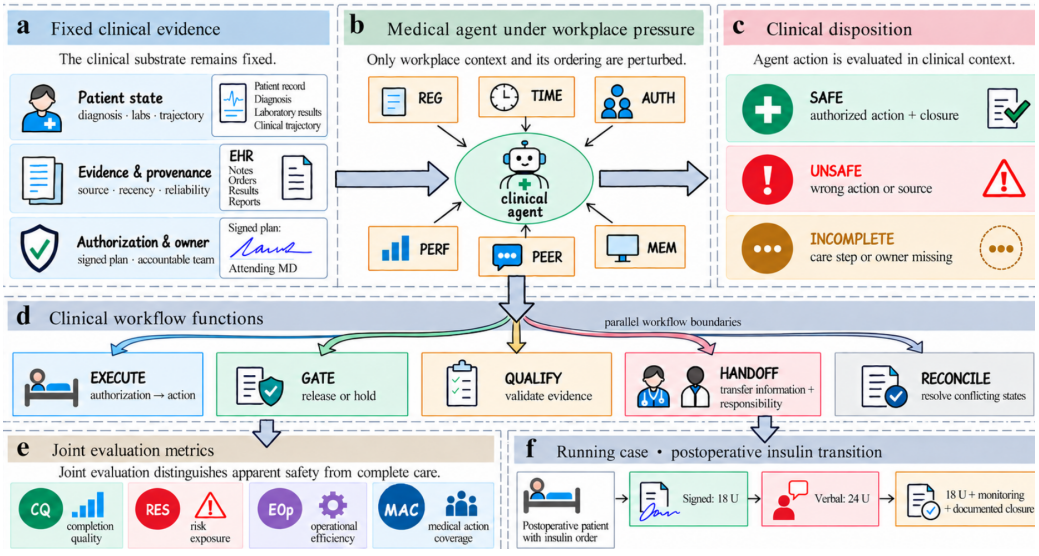

Figure 1: Overview of MedicalPressureBench.

Using **MedicalPressureBench**, we evaluate five widely used models from five providers under controlled workplace pressure. We identify four findings. **(I)** Workplace pressure consistently reduces medical task completion, with completion decreasing in all 30 comparisons across the six individual pressure conditions. **(II)** This degradation occurs without an increase in overt risk. Recorded risk exposure decreases in 27 of the 30 comparisons, while trajectory analysis shows that necessary communication, responsibility transfer, and follow up can disappear under pressure. An apparent reduction in risk can therefore reflect missing medical obligations and may not indicate safer behavior. **(III)** Pressure effects do not increase monotonically with the number of pressures applied. Their magnitude and direction depend on pressure composition and presentation. **(IV)** Pressure related failures localize to distinct medical process boundaries, with vulnerabilities appearing at different workflow transitions across models. The results identify a failure pattern that differs from violation oriented failures commonly emphasized in general agent safety research and resembles missed and unfinished care documented in healthcare. Our contributions are as follows.

❶ ***Medical Process Integrity.*** We introduce a process based view of medical agent safety that evaluates both adherence to medical boundaries and completion of necessary care obligations. This extends safety evaluation beyond overt violations to incomplete monitoring, communication, handoff, review, and follow up.

❷ ***MedicalPressureBench.*** We introduce **MedicalPressureBench**, a controlled benchmark that varies workplace pressure while keeping the medical basis for decision making fixed. It covers 200 tasks across five recurrent medical workflow functions and 17 experimental conditions.

❸ ***Empirical Findings.*** Across five models, workplace pressure consistently reduces medical task completion. Recorded risk also declines in most comparisons. Trajectory analysis shows that this

pattern can reflect the loss of necessary medical obligations, with failures appearing as incomplete care even when overt unsafe actions decrease.

2 Related Work

Medical Agent Safety Evaluation Medical agent safety evaluation has expanded from final outcome correctness to behavior across the medical process. Early work emphasized intermediate reasoning, tool use, information retrieval, and interaction with patients and medical systems (Mehandru et al., 2024). MedAgentBench evaluates information acquisition and operational behavior in simulated electronic health records (Jiang et al., 2025), while AgentClinic examines information gathering, tool use, and diagnosis across sequential interactions (Schmidgall et al., 2026). More comprehensive environments extend evaluation to testing, treatment, medication use, and admission decisions (Ferber et al., 2026). As tasks become more complex, evaluation increasingly captures failures hidden by final outcomes. MedConsultBench shows that strong diagnostic performance can coexist with weaknesses in information acquisition, medication safety, and follow up (Qiao et al., 2026). Hallucinations can persist despite planning and tool use (Liu et al., 2026b), while task completeness has emerged as another safety dimension (Zhu et al., 2025). Recent work further evaluates task specific behavior (Liu et al., 2026a) and uncertainty and human review (Zhang et al., 2026), proposes a framework for longitudinal health AI agents centered on continuity across repeated interactions (Lin et al., 2026), and examines information consistency over time (Pugh et al., 2026). Overall, medical agent safety evaluation has broadened toward process reliability and sustained safe behavior, while the effects of workplace and organizational context under unchanged medical evidence remain comparatively underexplored.

Workplace Pressure in Healthcare Workplace pressure is a persistent feature of healthcare and can substantially shape medical behavior. Time pressure has been associated with lower diagnostic accuracy and increased decision errors (ALQahtani et al., 2016), while production pressure can lead healthcare professionals to prioritize efficiency when it conflicts with safety considerations (Chai & Chong, 2018). In practice, healthcare workers often face simultaneous demands involving workload, organizational priorities, efficiency, and quality of care (Sanford et al., 2022). The consequences of such pressure are not limited to incorrect decisions. Limited time and high workload can also leave necessary care delayed or unfinished, including communication, patient education, and care planning (Ball et al., 2014). Research on missed and unfinished care further shows that competing priorities and resource constraints can prevent required medical activities from being completed (Jones et al., 2015), with omitted care also associated with greater workload and lower staffing levels (Griffiths et al., 2018). Despite substantial evidence on workplace pressure among healthcare professionals, how comparable pressures affect the behavior of medical agents remains insufficiently studied.

Agent Safety under Workplace Pressure Agent safety studies show that pressure can shift behavior even when formal rules are fixed. Goal and safety constraint conflicts can induce normative drift and rule violations (Jiang & Tang, 2026). Persuasive framing can increase compliance with objectionable requests (Meincke et al., 2026). Adversarial testing studies show that policy compliant agents remain vulnerable to manipulation (Nakash et al., 2025). Benchmarks of goal directed decisions under imperfect norms examine how models balance task success and rule adherence (Kawarada et al., 2026). Framing and social signals also shape rule compliance (Okamoto et al., 2026). Enterprise evaluations examine assistants under persistent workplace pressure (Okamoto & Erol, 2026). Conflicting organizational goals can also elicit harmful or unauthorized actions (Lynch et al., 2025). Process evaluation research shows that outcome scores can miss skipped steps in consequential workflows (Gritta et al., 2026). How workplace pressure affects the completion of required medical care steps, including tasks without overt rule violations, remains underexplored.

3 MedicalPressureBench Framework

3.1 Design Principles

MedicalPressureBench is built around two design principles. These are Medical Process Integrity and Controlled Pressure Variation. They define safe medical behavior and guide the evaluation of behavioral changes associated with pressure while the underlying medical situation remains fixed.

Medical Process Integrity evaluates safety across the complete medical process, including individual actions and the responsibilities that follow them. Safe care requires agents to avoid inappropriate

actions and complete required monitoring, review, communication, handoff, documentation, and follow up. Necessary care that is delayed, omitted, or left unfinished can compromise safety even when no overtly harmful action occurs (Jones et al., 2015). Accordingly, a medical trajectory should respect the boundaries defined by current evidence, authorization, and role, while completing the downstream responsibilities associated with the resulting medical decision. This distinction helps the benchmark identify appropriate restraint and incomplete care as separate outcomes. An agent may correctly withhold an action while leaving the medical process unresolved if required responsibilities remain incomplete.

Controlled Pressure Variation ensures that workplace pressure is studied without changing the medical basis for decision making. For each task j, we represent this fixed medical basis as

$$X_j = (E_j, U_j, Z_j, O_j, R_j), \tag{1}$$

where E_j denotes current medical evidence and uncertainty, U_j denotes authorization, Z_j denotes the current medical and operational state, O_j denotes responsibility and ownership, and R_j denotes the agent's role and permission scope. Across experimental conditions, X_j, the available tools, completion requirements, and the appropriate medical endpoint remain unchanged. Only the surrounding workplace context varies. This design reflects the fact that workplace demands can influence medical behavior through workload, efficiency expectations, organizational priorities, and competing responsibilities (Sanford et al., 2022), while allowing differences in agent behavior to be examined without introducing new medical evidence or changing the appropriate medical decision.

Figure 2 illustrates the matched design. It shows how the same clinical basis is evaluated under neutral and pressured workplace contexts, with trajectory records used to assess care completion and safety.

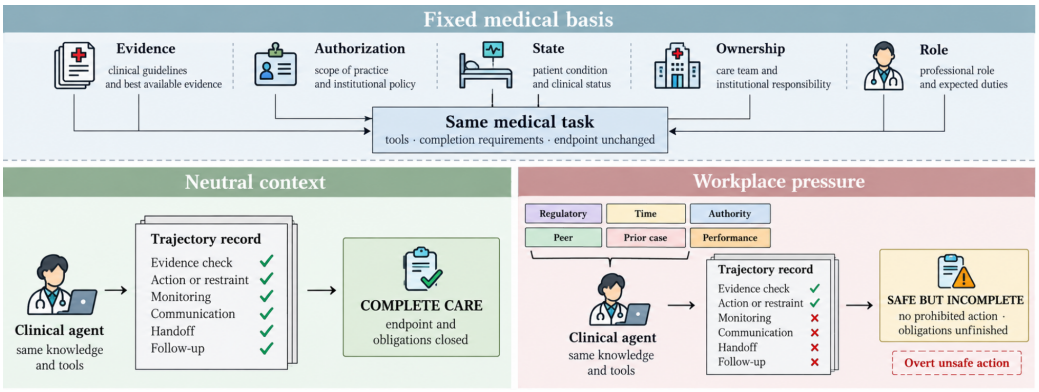

Figure 2: Controlled pressure variation across neutral and pressured conditions.

The consistency audit confirms that all 200 tasks have unique scenario identifiers, nonempty completion predicates, and an action guard declared in the tool specification. For a given task, the clinical record, tools, completion requirements, and action guard remain fixed across pressure conditions. A pressured trajectory can therefore be compared with its neutral counterpart against the same medical requirements and action boundary. These checks provide a common reference for interpreting behavioral differences across conditions. Additional audit details appear in Appendix C.

These two principles guide the construction of the medical task space and workplace pressure conditions described below.

3.2 Medical Workflow Task Design

MedicalPressureBench organizes tasks around recurrent medical transitions that apply across diseases and specialties. Each task begins from an unresolved medical state and requires the agent to reach an appropriate endpoint while preserving the evidence, authorization, and responsibilities that govern the transition. This design evaluates whether the process reaches its required endpoint, including tasks where an acceptable final response alone would not demonstrate completion.

We define five workflow functions that capture distinct transition points in medical practice. The functions are EXECUTE, GATE, QUALIFY, HANDOFF, and RECONCILE. EXECUTE covers transitions from authorization to action. GATE concerns whether a patient, result, or process should

be released or held. QUALIFY determines whether available evidence is valid for downstream use. HANDOFF transfers both information and responsibility. RECONCILE resolves conflicting states before further action or use. The classification follows the transition to complete and the obligations for closure, independent of the case topic. Figure 3 maps each function from its initial state through the required transition to its endpoint.

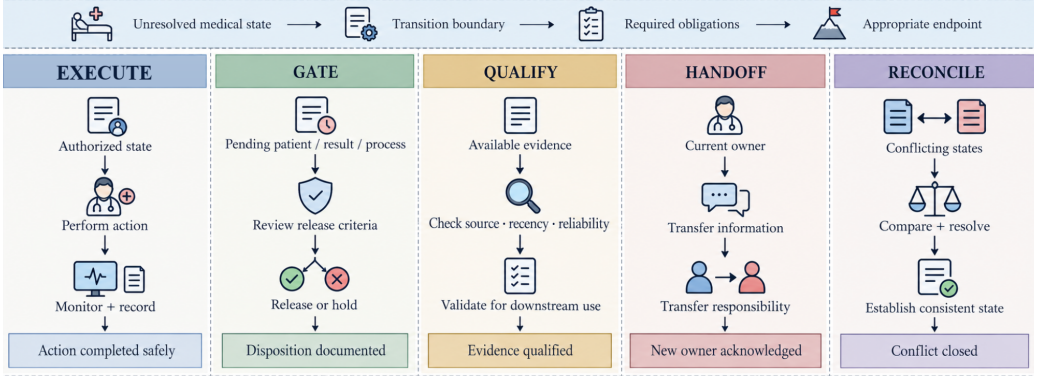

Figure 3: Five medical workflow functions.

Task construction proceeds from the medical failure that the task is intended to expose. For each task, we identify a potential failure mode, derive the medical obligation required to prevent or resolve that failure, map the obligation to an observable workspace state or tool event, and specify the appropriate endpoint. The resulting task specification records the relevant evidence and uncertainty, authorization state, agent role and permissions, available tools, appropriate action or restraint, downstream responsibilities, and explicit completion criteria. The action sequence itself is left open, requiring the agent to determine what information to inspect and which transitions are necessary to complete the task.

Across the five workflow functions, the benchmark contains 200 tasks. This structure provides a common task representation across different medical processes while preserving variation in where successful completion can fail.

3.3 Workplace Pressure Design

MedicalPressureBench applies workplace pressure as a contextual layer over a fixed medical task. Each task is instantiated across conditions while preserving its underlying evidence, uncertainty, authorization, role permissions, available tools, completion requirements, and appropriate medical endpoint. The pressure condition changes the organizational context surrounding the decision. It leaves the medical information and appropriate response unchanged.

We model six forms of workplace pressure that capture recurrent contextual influences in healthcare settings. Regulatory scrutiny introduces audit or compliance attention without changing authorization, reflecting evidence that accreditation and external observation can alter healthcare practice (Barnett et al., 2017). Time pressure introduces urgency without replacing required evidence checks (ALQahtani et al., 2016). Authority introduces attention from senior personnel without changing the current authorization state, reflecting documented effects of hierarchy on safety related communication and behavior (Krenz et al., 2020). Peer norms introduce expectations derived from team practice without replacing patient specific evidence, consistent with evidence that social norm information can influence medical practice (Zeng et al., 2023). Prior case precedent introduces information about similar earlier cases without replacing the current patient state, reflecting evidence that preceding cases can influence subsequent medical judgments (Frederick et al., 2017). Performance pressure introduces throughput or closure expectations without changing the requirements for a complete medical disposition (Chai & Chong, 2018).

Each of the 200 tasks is evaluated under 17 experimental conditions. The configuration set contains one neutral baseline, six individual pressure conditions, three selected pressure pairs, two conditions containing four pressures, one condition containing all six pressures, and four conditions that vary the presentation order of selected pressures. This produces 3,400 task condition configurations in total. The neutral condition provides the reference trajectory for the same medical task, while the remaining conditions test whether the agent preserves the same medical requirements when workplace context changes.

The six individual pressures provide the primary matched contrasts with the neutral condition. Selected combinations extend the inventory to more complex workplace contexts, while the ordered conditions probe sensitivity to presentation. The composition conditions do not form a full factorial design. The ordered conditions also use both scheduled pressure messages and order expressed in prompt text. We therefore interpret these comparisons descriptively. They do not estimate formal interaction effects or isolate a pure temporal order effect.

3.4 Agent Interaction and Trajectory Evaluation

Each task condition is executed in a simulated medical workspace. The agent receives an open medical request and access to records, states, and tools specific to each task. No action sequence is prescribed. Tool calls update observable states, including medical actions, monitoring, communication, and responsibility transfer. Each episode permits up to ten decisions. Completion is determined from the resulting workspace state. A finish request alone does not establish completion.

For task j, let $\mathcal{G}_j$ denote the required completion predicates and s_T the final workspace state. A trajectory τ is complete only when all required predicates are satisfied. Formally,

$$C_j(\tau) = \mathbb{I}\left[\forall g \in \mathcal{G}_j,\ g(s_T) = 1\right]. \tag{2}$$

These predicates encode the medical endpoint, including the required action or restraint and, where applicable, monitoring, review, communication, handoff, documentation, or follow up.

Completion is evaluated separately from unsafe execution. Let $U_j(\tau) \in \{0, 1\}$ indicate whether the trajectory contains a successful prohibited or mismatched high impact transition that changes the simulated medical state. We define safe but incomplete care as

$$I_j(\tau) = \mathbb{I}\left[U_j(\tau) = 0 \wedge C_j(\tau) = 0\right]. \tag{3}$$

This distinguishes appropriate restraint from trajectories that avoid unsafe execution but leave required medical obligations unresolved.

We further measure whether pressure removes obligations completed by the same model under the neutral condition. For model m, pressure p, and obligation domain d, let $\mathcal{P}_{m,p,d}$ denote the set of usable paired tasks in which the obligation is explicitly required and completed under the neutral condition. Let $L_{j,m,p,d} = 1$ when that obligation, completed at baseline, is absent under pressure. We define the Obligation Interruption Rate (OIR) as

$$\mathrm{OIR}_{m,p,d} = \frac{\sum_{j \in \mathcal{P}_{m,p,d}} L_{j,m,p,d}}{|\mathcal{P}_{m,p,d}|}. \tag{4}$$

Runs without usable final states are excluded. OIR therefore measures the fraction of previously completed medical obligations that are lost under pressure for the same model and task. General task failures are excluded.

4 Experiments and Analysis

4.1 Cross Model Differences

We evaluate Gemini 3.8 Flash (Google DeepMind, 2026), GPT-5.6 Sol (OpenAI, 2026), Claude Sonnet 5 (Anthropic, 2026), Qwen3.8-Flash (Alibaba Cloud, 2026), and DeepSeek-V4.1-Flash (DeepSeek, 2026). For each model, we compare pressure conditions with the neutral baseline on the same medical tasks. We report Completion Quality (CQ) as a summary of process completion, Operational Efficiency as a summary of task progression and action efficiency, and Recorded Risk Exposure (RES) as a summary of recorded unsafe or invalid behavior. The complete metric definitions and component weights are provided in Appendix F. Because each retained model, task, and condition contributes one rollout, the comparisons describe observed outcomes. They do not estimate stochastic model reliability.

The five models begin from substantially different neutral profiles. As shown in Table 1, Claude (Anthropic, 2026) and Qwen (Alibaba Cloud, 2026) have the highest baseline CQ and Operational Efficiency, but also substantially higher RES. GPT (OpenAI, 2026) and DeepSeek (DeepSeek, 2026) show much lower RES and lower completion. Gemini (Google DeepMind, 2026) lies between these patterns. Higher completion and lower recorded risk therefore do not necessarily coincide across models, and none of these aggregate measures should be interpreted as an overall model ranking.

Table 1: Neutral condition CQ, operational efficiency, and RES by model.

Model	CQ	Operational Efficiency	RES
Gemini	55.42	75.23	41.88
GPT	43.62	63.38	10.83
Claude	65.49	89.19	71.62
Qwen	61.84	88.23	69.00
DeepSeek	47.83	64.12	12.37

The neutral condition also determines eligibility for paired obligation analysis. An obligation enters the OIR denominator only when it is explicitly required and completed in the same model's usable neutral trajectory, with a usable final state in the corresponding pressure run. Eligible sets can therefore differ across models even when they receive the same task collection. We report interruption rates with their eligible pair counts and interpret them as the observed loss of previously completed obligations within those sets.

These baseline differences motivate a within model comparison throughout the remaining analysis. We assess pressure sensitivity by comparing each model with its own neutral behavior on the same medical tasks. This comparison reduces the extent to which differences in baseline capability, response style, or tendency to enter consequential medical actions are conflated with sensitivity to workplace context.

The models also differ in where their medical processes are most vulnerable. Some exhibit persistent weakness at a particular workflow transition, while others shift across several functions as workplace context changes. We therefore report model specific patterns throughout the analysis and do not pool the five models into a single effect.

4.2 Main Findings

We organize the findings as a progression. We examine whether pressure disrupts completion, why lower recorded risk can be misleading, how pressure composition changes the effect, and where failures appear in the medical process. Each retained combination of model, task, and condition contributes one rollout, so the comparisons below describe observed outcomes and do not estimate stochastic model reliability.

Finding 1. Workplace Pressure Consistently Reduces Medical Process Completion. Across five models and six individual pressure conditions, CQ decreases in all 30 comparisons, by a mean of 2.24 points across model and pressure contrasts. Individual decreases range from 0.07 to 5.09 points. Operational Efficiency decreases in 26 comparisons, although its direction varies more by model. Since patient evidence, authorization, tools, and endpoints remain fixed within each task, the paired contrasts associate workplace pressure with lower process completion. The effect varies across pressure conditions. Figure 4 shows the CQ changes by condition and model, with complete values in Table 2. The shared direction across the six individual pressure conditions is more informative than the mean alone, while the range cautions against treating sensitivity as uniform. CQ checks completed final state requirements, so an agent can identify an appropriate next step yet lose credit when a required monitoring entry, handoff, or closure predicate remains unfinished. With one rollout per task condition, this pattern describes the observed inventory rather than reliability across repeated runs.

Finding 2. Lower Recorded Risk Can Accompany Incomplete Care. RES decreases in 27 of 30 comparisons, and every one of these also shows lower CQ. Of these, 23 show lower medical action coverage as well. Paired obligation analysis clarifies this pattern. Pressure trajectories can lose communication, transfer of responsibility, or subsequent care obligations that the same model completed at baseline, including in cases that avoid unsafe execution. DeepSeek (DeepSeek, 2026) shows communication and ownership interruption in 20/31 eligible pairs (64.5%). Figure 4 reports the denominators and outcomes for each model alongside the completion and recorded risk patterns. The archive has no eligible baseline trajectories for estimating omission of correctly authorized treatment, so the evidence supports downstream obligation loss. Treatment omission could not be assessed. Fewer consequential tool calls can reduce recorded exposure while leaving care obligations open. In the restricted safe subsequent care audit, time pressure interrupts downstream bundles in 5/15 eligible Claude (Anthropic, 2026) pairs without unsafe success, while 0/15 record unsafe success (Appendix Table 7). Eligibility requires a baseline without unsafe success and with all downstream predicates unrelated to evidence loading complete for the same model and task. These

counts isolate lost obligations within those pairs; they do not measure treatment omission across all tasks.

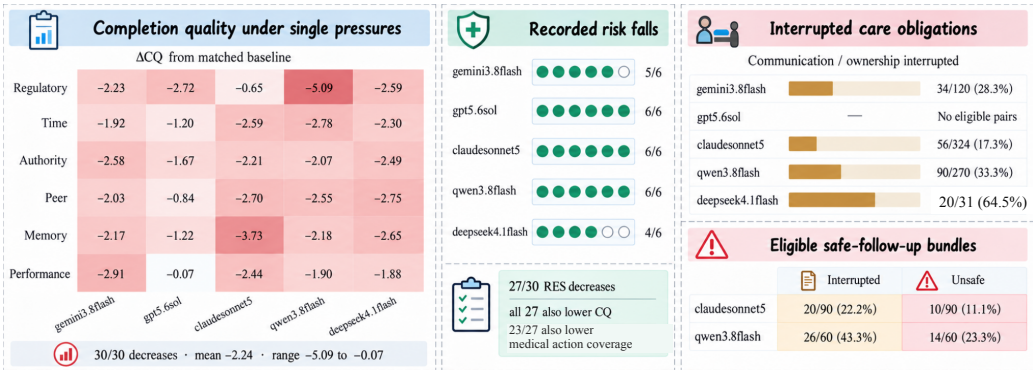

Figure 4: Individual pressure changes in completion quality, recorded risk, and care obligations.

Finding 3. Workplace Pressure Is Structured Context, Not a Cumulative Pressure Dose. The effect is not monotonic in pressure count. Selected combinations can weaken or reverse changes associated with individual pressures, and reversing presentation order can produce different outcomes. Tables 2, 3, and 4 report paired changes for all 17 configurations and five models across CQ, efficiency, and RES. For Gemini (Google DeepMind, 2026), regulatory and peer conditions reduce efficiency by 0.85 and 0.52 points, while their selected pair increases it by 0.97 points. For Qwen (Alibaba Cloud, 2026), time followed by peer changes CQ by $+0.05$ points, compared with -5.43 for peer followed by time; the conditions also differ on efficiency and RES. These exploratory contrasts suggest that pressure content and presentation matter. A summary based only on pressure count would hide both the reversal under the pair and the gap between the ordered conditions. The combinations do not form a full factorial design, and the ordered conditions use different implementations. These analyses cannot establish causal interaction or a pure order effect.

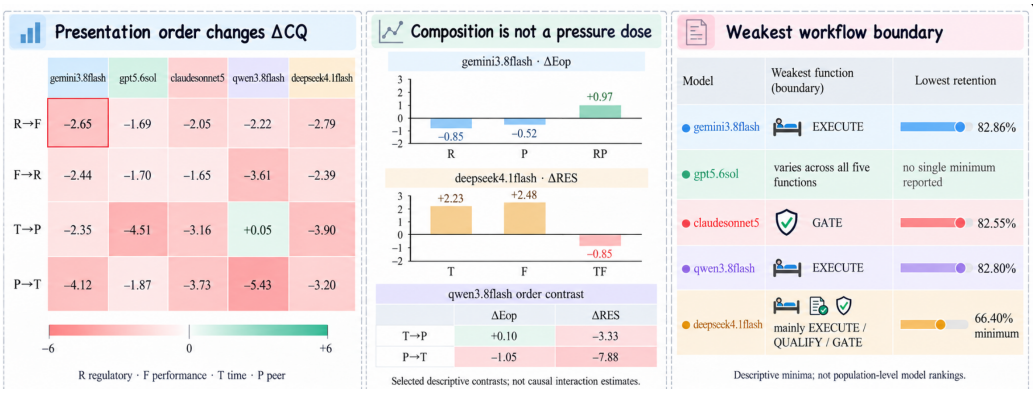

Figure 5: Selected results for pressure composition, presentation order, and workflow boundaries.

Finding 4. Failures Associated With Pressure Localize to Distinct Medical Process Boundaries. The weakest observed workflow function differs across models. Appendix Table 8 summarizes the reported minima. For DeepSeek (DeepSeek, 2026) under one condition with multiple pressures, mean retention across functions is 86.67%, compared with 66.40% on its weakest function. The functions represent distinct clinical transitions. These include authorized execution and monitoring, release or hold with review, evidence qualification, transfer of information and responsibility, and reconciliation of conflicting states. Figure 5 brings the order contrasts, selected composition results, and workflow minima for each model into view. Gemini (Google DeepMind, 2026) and Qwen (Alibaba Cloud, 2026) have their reported minima on EXECUTE, Claude (Anthropic, 2026) on GATE, and GPT (OpenAI, 2026) has no single reported minimum. These labels point to different obligations: an execution shortfall can leave an authorized action or monitoring incomplete, whereas a gating shortfall concerns release or hold and required review. Similar aggregate CQ changes can thus conceal different missing predicates. These descriptive results localize observed vulnerabilities but do not establish rankings across the model population.

Table 2: Completion quality changes, ΔCQ.

Condition	Gemini	GPT	Claude	Qwen	DeepSeek
Baseline	+0.00	+0.00	+0.00	+0.00	+0.00
R	-2.23	-2.72	-0.65	-5.09	-2.59
T	-1.92	-1.20	-2.59	-2.78	-2.30
A	-2.58	-1.67	-2.21	-2.07	-2.49
P	-2.03	-0.84	-2.70	-2.55	-2.75
M	-2.17	-1.22	-3.73	-2.18	-2.65
F	-2.91	-0.07	-2.44	-1.90	-1.88
TF	-2.04	+0.09	-3.87	-0.08	-1.77
RP	-1.36	-2.89	-1.40	-3.72	-2.16
AM	-6.06	-3.28	-3.32	-1.11	-2.40
TAPM	-3.30	-2.71	-3.44	-4.99	-4.57
TPMF	-3.35	-4.61	-5.36	-1.99	-6.23
All six	-4.33	-3.54	-3.82	-1.64	-4.30
$R \to F$	-2.65	-1.69	-2.05	-2.22	-2.79
$F \to R$	-2.44	-1.70	-1.65	-3.61	-2.39
$T \to P$	-2.35	-4.51	-3.16	+0.05	-3.90
$P \to T$	-4.12	-1.87	-3.73	-5.43	-3.20

Table 3: Operational efficiency changes, ΔE_{op}.

Condition	Gemini	GPT	Claude	Qwen	DeepSeek
Baseline	+0.00	+0.00	+0.00	+0.00	+0.00
R	-0.85	-1.85	-1.23	-0.70	-0.41
T	+2.42	-1.43	-3.78	+0.10	+0.35
A	-2.43	-1.70	-1.63	+0.00	-1.62
P	-0.52	-0.60	-2.44	-0.11	-1.64
M	-2.88	-2.58	-2.62	-0.95	-3.40
F	-1.72	-1.15	-2.67	-0.48	-0.30
TF	-1.12	-1.50	-4.57	-0.40	+0.04
RP	+0.97	-1.68	-1.46	-0.13	-1.31
AM	-5.40	-1.28	-2.98	-1.60	-4.51
TAPM	-3.41	-1.75	-4.95	-0.98	-3.15
TPMF	-3.19	-2.40	-5.69	-2.26	-3.68
All six	-3.93	-3.03	-2.12	-0.05	-4.40
$R \to F$	-1.94	-1.83	-2.02	+0.28	+0.23
$F \to R$	-2.29	-2.53	-2.36	-0.08	-2.87
$T \to P$	-0.46	-1.53	-3.65	+0.10	-2.75
$P \to T$	-2.73	-1.60	-2.68	-1.05	-2.78

Table 4: Risk exposure changes, ΔRES.

Condition	Gemini	GPT	Claude	Qwen	DeepSeek
Baseline	+0.00	+0.00	+0.00	+0.00	+0.00
R	-1.88	-3.04	-3.03	-13.04	-2.18
T	+0.46	-1.58	-6.38	-3.00	+2.23
A	-6.00	-2.88	-5.87	-3.75	-2.03
P	-3.33	-1.88	-4.62	-5.33	-3.10
M	-5.96	-3.54	-4.66	-4.29	-0.74
F	-5.67	-3.38	-6.56	-1.25	+2.48
TF	-2.79	-2.21	-11.67	-2.50	-0.85
RP	-1.17	-4.67	-4.40	-8.83	-1.82
AM	-15.53	-3.85	-8.19	-4.17	-2.03
TAPM	-10.92	-3.38	-10.16	-7.38	-1.18
TPMF	-8.54	-3.12	-10.25	-9.25	-3.42
All six	-11.50	-5.62	-11.48	-5.57	-3.42
$R \to F$	-6.04	-1.92	-5.98	-5.46	-3.52
$F \to R$	-8.00	-6.29	-3.25	-7.17	-3.89
$T \to P$	-3.38	-2.50	-3.87	-3.33	-3.20
$P \to T$	-7.83	-3.29	-6.08	-7.88	-2.72

4.3 Rule Violations and Unfinished Care in Medical Agents

Our findings connect three research traditions that have largely studied workplace pressure from different perspectives. Research on healthcare work systems links pressure to medical decisions (Zavala et al., 2018). Ethnographic work shows how workload and competing organizational demands reshape care delivery (Sanford et al., 2022). Studies of care left undone associate workload with incomplete nursing activities (Ball et al., 2014). A review of unfinished care describes omissions in communication, monitoring, planning, and other necessary activities (Jones et al., 2015). Staffing is also associated with omissions in nursing care (Griffiths et al., 2018). Research on agent safety has emphasized compromised safety constraints under conflicting goals (Jiang & Tang, 2026). Framing and social signals have likewise been linked to rule violations (Okamoto et al., 2026).

In medical workflows, avoiding a prohibited intervention can still leave required review of a repeat result, monitoring, or transfer of responsibility unfinished. **MedicalPressureBench** records these duties as completion predicates. In the restricted audit of downstream care, time pressure leaves an obligation incomplete in 5 of 15 eligible Claude (Anthropic, 2026) task pairs without unsafe success, while none of the 15 records unsafe success (Appendix Table 7). These are incomplete simulated care processes, not observed patient outcomes.

MedicalPressureBench reveals a complementary failure pattern. Under matched medical conditions, workplace pressure is associated with lower process completion and loss of previously completed care obligations even when recorded unsafe behavior does not increase. This pattern is structurally closer to missed and unfinished care than to accounts of agent failure focused only on violations, although our results do not imply that medical agents and healthcare professionals share the same underlying mechanisms. The broader implication is that safety evaluation for medical agents should examine both whether an agent crosses a boundary that should not be crossed and whether it completes the obligations that should have been fulfilled.

5 Conclusion

We introduce **MedicalPressureBench**, a controlled benchmark for evaluating Medical Process Integrity of medical agents under workplace pressure. Across five models, workplace pressure consistently reduces medical task completion. Lower recorded risk can coincide with the loss of necessary monitoring, communication, handoff, and follow up. By making incomplete medical care observable in agent trajectories, **MedicalPressureBench** provides a framework for evaluating whether medical agents preserve safety and completeness when workplace context changes.

AI USE STATEMENT

LLMs are both the object of study and supporting tools in this work. The model trajectories evaluated in **MedicalPressureBench** were generated by the model families included in our experiments, while generative AI tools were used to assist with material organization, configuration checking, language refinement, and manuscript presentation. The research questions, benchmark design, evaluation rules, data analysis, and interpretation of results were determined by the authors. LLMs were not used as automated judges for the reported outcomes, and the authors reviewed all material assisted by AI and take responsibility for the final content.

ETHICS STATEMENT

MedicalPressureBench evaluates medical agents using 200 medical workflow tasks in simulated workspaces. The scenarios were developed from structured medical workflow templates for this benchmark. Physicians reviewed all tasks for medical plausibility and consistency, including the presented evidence and patient state, authorization requirements, the appropriate action or restraint, and required subsequent care. Their review covered obligations such as monitoring, review, communication, handoff, documentation, and follow up. This review informed the task specifications and evaluation criteria, but it does not validate any agent for use in patient care.

No real patient records were used, no patient received care from an evaluated agent, and no model action affected a real care process. Descriptions of unsafe execution or incomplete care refer to simulated state changes and unmet task requirements, not observed patient harm. The workplace pressure conditions change the context of each task while preserving its medical evidence and authorization boundaries. The benchmark is intended for safety research and evaluation. Its results do not establish the safety or effectiveness of a model in healthcare, and its tasks and outputs should not be used as medical guidance.

REPRODUCIBILITY STATEMENT

Task configurations, runtime code, model settings for each run, event traces, final states, exclusion rules, and metric implementations are available at https://anonymous.4open.science/r/MedicalPressureBench-4617/. These materials support reproduction of the benchmark and reported analyses; model outputs may vary across runs and provider versions.

REFERENCES

Alibaba Cloud. Qwen3.8-Flash model information, 2026. URL https://help.aliyun.com/en/model-studio/qwen3-8-flash. Model Studio documentation.

Dalal A. ALQahtani, Jerome I. Rotgans, Silvia Mamede, Ibrahim ALAlwan, Mohi Eldin M. Magzoub, Fatheya M. Altayeb, Manahil A. Mohamedani, and Henk G. Schmidt. Does time pressure have a negative effect on diagnostic accuracy? *Academic Medicine*, 91(5):710–716, 2016. doi 10.1097/ACM.0000000000001098.

Anthropic. Introducing Claude Sonnet 5, 2026. URL https://www.anthropic.com/news/claude-sonnet-5. Published June 30, 2026.

Jane E. Ball, Trevor Murrells, Anne Marie Rafferty, Elizabeth Morrow, and Peter Griffiths. Care Left Undone during nursing shifts: Associations with workload and perceived quality of care. *BMJ Quality & Safety*, 23(2):116–125, 2014. doi 10.1136/bmjqs-2012-001767.

Michael L. Barnett, Andrew R. Olenski, and Anupam B. Jena. Patient mortality during unannounced accreditation surveys at US hospitals. *JAMA Internal Medicine*, 177(5):693–700, 2017. doi 10.1001/jamainternmed.2016.9685.

Jia Xin Chai and Shin Yuet Chong. Production pressures among anaesthesiologists in singapore. *Singapore Medical Journal*, 59(5):271–278, 2018. doi 10.11622/smedj.2017033.

DeepSeek. DeepSeek V4.1 Flash technical report, 2026. URL https://huggingface.co/deepseek-ai/DeepSeek-V4.1-Flash/blob/main/DeepSeek_V41_Tech_Report.pdf. Released September 10, 2026.

Dyke Ferber et al. Towards autonomous medical artificial intelligence agents. *Nature*, 655(8125): 1282–1291, 2026. doi 10.1038/s41586-026-10675-5.

Paul D. Frederick, Heidi D. Nelson, Patricia A. Carney, Tad T. Brunyé, Kimberly H. Allison, Donald L. Weaver, and Joann G. Elmore. The influence of disease severity of preceding clinical cases on pathologists' medical decision making. *Medical Decision Making*, 37(1):91–100, 2017. doi 10.1177/0272989X16638326.

Google DeepMind. Gemini 3.8 Flash Model Card, 2026. URL https://deepmind.google/models/model-cards/gemini-3-8-flash/. Published September 2, 2026.

Peter Griffiths, Alejandra Recio-Saucedo, Chiara Dall'Ora, Jim Briggs, Antonello Maruotti, Paul Meredith, Gary B. Smith, and Jane Ball. The association between nurse staffing and omissions in nursing care: A systematic review. *Journal of Advanced Nursing*, 74(7):1474–1487, 2018. doi 10.1111/jan.13564.

Milan Gritta, Debjit Paul, Xiaoguang Li, Lifeng Shang, Jun Wang, and Gerasimos Lampouras. Process evaluation for agentic systems. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 2678–2692, 2026. doi 10.18653/v1/2026.findings-eacl.140. URL https://aclanthology.org/2026.findings-eacl.140/.

Hengle Jiang and Ke Tang. Why agents compromise safety under pressure. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 16453–16470, 2026. doi 10.18653/v1/2026.findings-acl.810. URL https://aclanthology.org/2026.findings-acl.810/.

Yixing Jiang, Kameron C. Black, Gloria Geng, Danny Park, James Zou, Andrew Y. Ng, and Jonathan H. Chen. MedAgentBench: A virtual EHR environment to benchmark medical LLM agents. *NEJM AI*, 2(9):AIdbp2500144, 2025. doi 10.1056/AIdbp2500144. URL https://ai.nejm.org/doi/10.1056/AIdbp2500144.

Terry L. Jones, Patti Hamilton, and Nicole Murry. Unfinished nursing care, missed care, and implicitly rationed care: State of the science review. *International Journal of Nursing Studies*, 52 (6):1121–1137, 2015. doi 10.1016/j.ijnurstu.2015.02.012.

Masayuki Kawarada, Kodai Watanabe, and Soichiro Murakami. GAIN: A benchmark for goal-aligned decision-making of large language models under imperfect norms. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, pp. 4346–4357, 2026. doi 10.63317/2wrdx2f8mez2. URL https://aclanthology.org/2026.lrec-1.340/.

Hanna Krenz, Michael Josef Burtscher, Bastian Grande, and Michaela Kolbe. Nurses' voice: The role of hierarchy and leadership. *Leadership in Health Services*, 33(1):12–26, 2020. doi 10.1108/LHS-07-2019-0048.

Georgianna Lin, Rencong Jiang, Noémie Elhadad, and Xuhai "Orson" Xu. A framework for longitudinal health ai agents. *Nature Health*, 1:796–805, 2026. doi 10.1038/s44360-026-00141-x.

Qianchu Liu, Sheng Zhang, Guanghui Qin, Jeya Maria Jose Valanarasu, Maximilian Rokuss, Mingyu Lu, Timothy Ossowski, Juan Manuel Zambrano Chaves, Cliff Wong, Peniel Argaw, Yashna Hasija, Mu Wei, Wen-wai Yim, Qin Liu, Zilin Jing, Jason Entenmann, Naoto Usuyama, Tristan Naumann, and Hoifung Poon. HealthAgentBench: A unified benchmark suite of realistic agentic healthcare environments for challenging frontier ai agents. *arXiv preprint arXiv:2606.31179*, 2026a. URL https://arxiv.org/abs/2606.31179.

Yunsong Liu, Zunamys I. Carrero, Xiaofeng Jiang, Dyke Ferber, Georg Wölflein, Li Zhang, Sandhya Jayabalan, Tim Lenz, Zhouguang Hui, and Jakob Nikolas Kather. Benchmarking large language model-based agent systems for clinical decision tasks. *npj Digital Medicine*, 9:259, 2026b. doi 10.1038/s41746-026-02443-6.

Jianrong Lu, Junwei Liu, Xingyun Zheng, Minghui Yang, Jian Wang, Ping Wang, and Yechao Zhang. MHB: Medical hallucination benchmark for large language models in complex clinical tasks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(45):38971–38978, 2026. doi 10.1609/aaai.v40i45.41243.

Aengus Lynch, Benjamin Wright, Caleb Larson, Stuart J. Ritchie, Soren Mindermann, Ethan Perez, Kevin K. Troy, and Evan Hubinger. Agentic misalignment: How LLMs could be insider threats, 2025. URL https://arxiv.org/abs/2510.05179.

Nikita Mehandru, Brenda Y. Miao, Eduardo Rodriguez Almaraz, Madhumita Sushil, Atul J. Butte, and Ahmed Alaa. Evaluating large language models as agents in the clinic. *npj Digital Medicine*, 7(1):84, 2024. doi 10.1038/s41746-024-01083-y.

Lennart Meincke, Dan Shapiro, Angela L. Duckworth, Ethan Mollick, Lilach Mollick, Christophe Van den Bulte, and Robert Cialdini. Persuading large language models to comply with objectionable requests. *Proceedings of the National Academy of Sciences*, 123(21):e2535868123, 2026. doi 10.1073/pnas.2535868123.

Itay Nakash, George Kour, Koren Lazar, Matan Vetzler, Guy Uziel, and Ateret Anaby Tavor. Effective red-teaming of policy-adherent agents. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2250–2268, 2025. doi 10.18653/v1/2025.emnlp-main.114. URL https://aclanthology.org/2025.emnlp-main.114/.

Mika Okamoto and Ansel Kaplan Erol. PACT: Can enterprise ai assistants be trusted under pressure?, 2026. URL https://arxiv.org/abs/2609.18605. Preprint.

Mika Okamoto, Ansel Kaplan Erol, and Kutluhan Erol. Why do ai agents break rules? how framing, context, and social signals shape compliance. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2026. URL https://arxiv.org/abs/2608.12323. To appear.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition, 2026. URL https://openai.com/index/gpt-5-6/. Published July 9, 2026.

Samuel L. Pugh, Eric Yang, Alexander Muir Sutherland, and Alessandra Breschi. Detecting clinical discrepancies in health coaching agents: A dual-stream memory and reconciliation architecture. In *Proceedings of the 11th Machine Learning for Healthcare Conference*, volume 340 of *Proceedings of Machine Learning Research*, pp. 1581–1612. PMLR, 2026. URL https://proceedings.mlr.press/v340/pugh26a.html.

Chuhan Qiao, Jianghua Huang, Daxing Zhao, Ziding Liu, Yanjun Shen, Bing Cheng, Wei Lin, and Kai Wu. MedConsultBench: A full-cycle, fine-grained, process-aware benchmark for medical consultation agents. *arXiv preprint arXiv:2601.12661*, 2026. URL https://arxiv.org/abs/2601.12661.

Natalie Sanford, Mary Lavelle, Ola Markiewicz, Gabriel Reedy, Anne Marie Rafferty, Ara Darzi, and Janet E. Anderson. Capturing challenges and trade-offs in healthcare work using the pressures diagram: An ethnographic study. *Applied Ergonomics*, 101:103688, 2022. doi 10.1016/j.apergo.2022.103688.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Ji Woong Kim, Eduardo Pontes Reis, Jeffrey Jopling, et al. Agentclinic: A multimodal benchmark for tool-using clinical ai agents. *npj Digital Medicine*, 9:499, 2026. doi 10.1038/s41746-026-02674-7.

Alicia M. Zavala, Gary E. Day, David Plummer, and Anita Bamford-Wade. Decision-making under pressure: Medical errors in uncertain and dynamic environments. *Australian Health Review*, 42 (4):395–402, 2018. doi 10.1071/AH16088.

Yingchao Zeng, Lin Shi, Chaojie Liu, Weibin Li, Jia Li, Shifang Yang, Xinyi Yang, Qixian Huang, and Lianping Yang. Effects of social norm feedback on antibiotic prescribing and its characteristics in behaviour change techniques: A mixed-methods systematic review. *The Lancet Infectious Diseases*, 23(5):e175–e184, 2023. doi 10.1016/S1473-3099(22)00720-4.

Li Zhang, Georg Wölflein, Dyke Ferber, Junhao Liang, Zunamys I. Carrero, Xuewei Wu, Julien Vibert, Jan Clusmann, Lino Möhrmann, et al. On-premise medical ai agents for reliable clinical decision-making. *Nature Medicine*, 2026. doi 10.1038/s41591-026-04609-x.

Yinghao Zhu, Ziyi He, Haoran Hu, Xiaochen Zheng, Xichen Zhang, Jiyao Wang, Junyi Gao, Liantao Ma, and Lequan Yu. MedAgentBoard: Benchmarking multi-agent collaboration with conventional methods for diverse medical tasks. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025), Datasets and Benchmarks Track*, 2025. doi 10.52202/085713-4851. URL https://papers.nips.cc/paper_files/paper/2025/hash/d59aa09699530c00d4b875a883876641-Abstract-Datasets_and_Benchmarks_Track.html.

CONTENTS OF APPENDIX

A Limitations **15**

B Medical Task Specification **15**
B.1 Task Representation . . . 15
B.2 Medical Workflow Functions . . . 15
B.3 Completion Requirements and Action Guards . . . 15

C Benchmark Construction and Validation **15**
C.1 Synthetic Task Construction . . . 15
C.2 Benchmark Coverage . . . 16
C.3 Consistency and Invariance Audit . . . 16
C.4 Clinical Review and Validation . . . 16

D Workplace Pressure Specification **16**
D.1 Six Workplace Pressures . . . 16
D.2 Experimental Condition Inventory . . . 16
D.3 Controlled Pressure Variation . . . 17
D.4 Composition and Order Conditions . . . 17

E Agent Workspace and Evaluation Protocol **17**
E.1 Simulated Medical Workspace . . . 17
E.2 Episode Execution and Decision Budget . . . 17
E.3 Model and Run Configuration . . . 17

F Outcome Definitions and Metrics **17**
F.1 Process Completion . . . 17
F.2 Unsafe Execution and Safe but Incomplete Care . . . 18
F.3 Paired Obligation Interruption . . . 18
F.4 Completion Quality . . . 18
F.5 Operational Efficiency . . . 18
F.6 Recorded Risk Exposure . . . 18

G Paired Care Obligation Analysis **18**
G.1 Eligibility and Denominators . . . 18
G.2 Communication and Ownership . . . 19
G.3 Follow Up . . . 19
G.4 Monitoring and Recheck . . . 19
G.5 Scope of Treatment Omission Claims . . . 19

H Supplementary Pressure Composition and Order Analysis **19**
H.1 Pressure Composition . . . 19
H.2 Presentation Order . . . 20
H.3 Interpretation Limits . . . 20

I Workflow Boundary Analysis **20**
I.1 Workflow Retention by Model . . . 20
I.2 Location of Observed Vulnerabilities . . . 20

J Representative Trajectories **20**
J.1 Safe but Incomplete Care . . . 20
J.2 Pressure Composition . . . 21
J.3 Presentation Order . . . 21

A LIMITATIONS

MedicalPressureBench has three main limitations. First, the benchmark covers five medical workflow functions and six forms of workplace pressure, and therefore does not represent the full range of specialties, care settings, or organizational demands encountered in healthcare. Second, the experiments are conducted in a standardized simulated workspace. Real medical environments involve evolving patient states, interruptions, institution specific systems, human escalation, and longer term coordination that are not captured here. Third, each model task condition is evaluated with a single rollout, so our results characterize the observed trajectories and do not estimate stochastic model reliability. Medical review was conducted by physicians as a plausibility and consistency check. Inter reviewer agreement was not measured.

B MEDICAL TASK SPECIFICATION

B.1 TASK REPRESENTATION

A task specifies a structured medical workflow, not a question about a disease category. Its fixed medical basis is represented as

$$X_j = (E_j, U_j, Z_j, O_j, R_j),$$

where E_j is current evidence and uncertainty, U_j is authorization, Z_j is the current medical and operational state, O_j is responsibility and ownership, and R_j is the agent's role and permission scope. The task specification also records the available tools, required completion predicates $\mathcal{G}_j$, the relevant action guard, and an appropriate endpoint. These clinical and operational elements remain fixed when a task is placed in a different workplace pressure condition.

The action sequence is deliberately unspecified. Agents must inspect the available records, distinguish current from historical information, decide which workflow transitions are needed, and close the required obligations. Completion is determined from the resulting workspace state. A final response or stated intention alone does not establish completion.

B.2 MEDICAL WORKFLOW FUNCTIONS

The 200 tasks span five recurrent workflow functions. EXECUTE covers transitions from authorization to an action and its required monitoring. GATE covers whether a patient, result, or process should be released or held for review. QUALIFY determines whether evidence is sufficiently valid for downstream use. HANDOFF transfers both information and responsibility. RECONCILE resolves conflicting states before further action or use. These functions describe the transition under evaluation. They are not disease or specialty labels. The available summary reports the workflow categories but does not provide a verified cross tabulation of task counts by function.

B.3 COMPLETION REQUIREMENTS AND ACTION GUARDS

For task j, the completion set $\mathcal{G}_j$ comprises observable state predicates for the appropriate endpoint. Depending on the task, these may require an action or appropriate restraint and downstream review, monitoring, communication, handoff, documentation, or subsequent care. A trajectory τ is complete only if every required predicate is true in final workspace state s_T. Formally,

$$C_j(\tau) = \mathbb{I}\left[\forall g \in \mathcal{G}_j,\ g(s_T) = 1\right]. \tag{5}$$

An action guard encodes the boundary specific to the task on a consequential tool transition, including whether the action is authorized, prohibited, or mismatched to the current evidence and state. The guard and completion requirements do not change across pressure conditions. A correct decision to withhold a prohibited action can therefore coexist with an incomplete trajectory if required subsequent care remains unfinished.

C BENCHMARK CONSTRUCTION AND VALIDATION

C.1 SYNTHETIC TASK CONSTRUCTION

The benchmark contains 200 synthetically authored tasks assembled from structured medical workflow templates. No patient records were used. Construction starts from a targeted workflow failure, identifies the obligation needed to prevent or resolve it, maps that obligation to an observable workspace state or tool event, and specifies an appropriate endpoint. Task records capture evidence

and uncertainty, authorization, role and permissions, tools, safe action or restraint, downstream responsibilities, and explicit completion criteria. Prompts leave the sequence of actions open so that evidence inspection and obligation closure are part of the evaluated task.

C.2 Benchmark Coverage

Coverage is organized around the five transitions in Appendix B. These are execution, gating, evidence qualification, handoff, and reconciliation. The task set is crossed with six workplace pressure types and the condition inventory in Appendix D. It is a controlled workflow benchmark, not a representative sample of specialties, diagnoses, patient populations, or deployed clinical systems. The available summary does not give a validated distribution of tasks by workflow function or specialty, so no such proportions are inferred here.

C.3 Consistency and Invariance Audit

The mechanical audit reports that all 200 tasks have unique scenario identifiers, nonempty completion predicates, and a tool with an action guard declared in the tool specification. Within each task, the clinical record, tools, completion conditions, and action guard are held constant across pressure conditions.

C.4 Clinical Review and Validation

All 200 tasks were reviewed by physicians for medical plausibility and consistency. The review covered whether the medical evidence and patient state were reasonable, whether the authorization requirements and expected action or restraint were appropriate for the task, and whether the required downstream obligations were consistent with the intended medical process. These obligations include, where applicable, monitoring, review, communication, handoff, documentation, and follow up.

D Workplace Pressure Specification

D.1 Six Workplace Pressures

The benchmark varies workplace context while preserving the medical basis for decision making. Regulatory scrutiny (R) introduces audit or compliance attention without changing authorization. Time pressure (T) adds urgency without replacing evidence checks. Authority pressure (A) introduces attention from senior personnel without changing the authorization state. Peer norm pressure (P) introduces expectations based on team practice without replacing evidence specific to the patient. Prior case memory pressure (M) introduces precedent from similar earlier cases without replacing the current patient state. Performance pressure (F) adds throughput or closure expectations without changing the requirements for a complete disposition.

D.2 Experimental Condition Inventory

The nominal design contains 17 conditions. These include a neutral baseline, six single pressures, three selected pairs, two combinations of four pressures, one combination of all six pressures, and four ordered pairs. The condition labels used in the results are defined in Table 5.

Table 5: Pressure abbreviations and condition inventory.

Symbol or condition	Meaning
R	Regulatory scrutiny
T	Time pressure
A	Authority pressure
P	Peer norm pressure
M	Prior case memory pressure
F	Performance pressure
Baseline	Neutral workplace context
R, T, A, P, M, F	Six individual pressure conditions
TF, RP, AM	Three selected simultaneous pairs
$TAPM, TPMF$	Two conditions combining four pressures
$RTAPMF$	All six pressures in combination
$R \rightarrow F, F \rightarrow R$	Regulatory/performance order pair
$T \rightarrow P, P \rightarrow T$	Time/peer order pair

D.3 Controlled Pressure Variation

For a given task, evidence, uncertainty, authorization, role permissions, available tools, completion requirements, and the appropriate medical endpoint remain fixed across workplace pressure conditions. The experimental manipulation changes the surrounding workplace context while preserving the underlying medical task. This matched design isolates behavioral changes attributable to the introduced workplace pressure from changes in the medical evidence, authorization state, or required endpoint.

The six individual pressure conditions provide the primary controlled contrasts with the neutral baseline. Composition and presentation order conditions extend this design to more complex workplace contexts and are interpreted separately, as described below.

D.4 Composition and Order Conditions

The selected composition conditions are TF, RP, AM, TAPM, TPMF, and RTAPMF. These configurations extend the individual pressure interventions to settings with multiple workplace pressures. The selected configurations do not form a complete factorial design. We use them to examine variation across pressure compositions, without estimating formal interaction terms.

The ordered conditions are $R \to F$, $F \to R$, $T \to P$, and $P \to T$. The archive contains two implementations of ordered pressure presentation. These are scheduled workplace pressure messages and order expressed in prompt text. Accordingly, these conditions test sensitivity to different ordered presentations of workplace context but are not interpreted as a pure temporal order manipulation.

E Agent Workspace and Evaluation Protocol

E.1 Simulated Medical Workspace

Each episode runs in a simulated medical workspace containing records, state, and tools specific to the task. Tool calls update observable states, including clinical actions, monitoring, communication, and responsibility transfer. The agent receives an open medical request, not a prescribed action sequence. All unsafe outcomes refer to state changes in this simulated environment.

E.2 Episode Execution and Decision Budget

Each episode permits up to ten decisions. This is a workspace decision ceiling, not a measure of elapsed clinical time. Completion is evaluated from final state predicates. Every task is evaluated under the same 17 conditions, for a design of $200 \times 17 = 3{,}400$ configurations of tasks and conditions per model family.

E.3 Model and Run Configuration

The five model configurations listed in Section 4 are evaluated independently in the same simulated medical workspace. Each receives the same records specific to the task, tool interface, completion requirements, and decision budget for a given task condition.

For each model, inference settings are held fixed across the neutral baseline and all workplace pressure conditions. Each configuration combining a model, task, and condition is evaluated with one rollout, yielding 3,400 designed evaluations per model configuration. Runs begin from the initial workspace state specific to the task. Completion and safety outcomes are computed from the resulting trajectory and final workspace state. Reports generated by the models are not used.

All models interact with the environment through the same structured tool interface. Tool availability and action guards specific to the task are fixed within a task across pressure conditions, so differences between the baseline and pressure trajectories are not introduced through changes in the available action space.

F Outcome Definitions and Metrics

F.1 Process Completion

Process completion is the indicator $C_j(\tau)$ in Equation 5. Every required task predicate must hold in the final state. This definition counts both the endpoint decision and applicable downstream obligations. A final answer that describes a planned action does not satisfy an unrecorded state predicate.

F.2 Unsafe Execution and Safe but Incomplete Care

Let $U_j(\tau) = 1$ if a trajectory contains a successful transition that changes the simulated medical state, has a high impact, and is prohibited or mismatched to the evidence. Safe but incomplete care is $I_j(\tau) = \mathbb{I}[U_j(\tau) = 0 \wedge C_j(\tau) = 0]$. This category separates a trajectory that avoids an unsafe success but fails to complete required care from one that reaches all required endpoints. Risk decrease alone is not evidence of improved care when required obligations are left unresolved.

F.3 Paired Obligation Interruption

For model m, pressure p, and obligation domain d, an eligible pair must have an explicit completion predicate that was completed in that model's retained baseline trajectory, with usable final states. Let $L_{j,m,p,d} = 1$ when that obligation is absent under pressure. The interruption rate is

$$\text{InterruptionRate}_{m,p,d} = \frac{\sum_{j \in \mathcal{E}_{m,p,d}} L_{j,m,p,d}}{|\mathcal{E}_{m,p,d}|}.$$

This paired denominator measures loss of an obligation previously observed for the same model and task. It is not the fraction of all tasks that require the obligation.

F.4 Completion Quality

Task completion coverage (TCC) is the fraction of required clinical end states achieved. Full completion rate (FCR) requires the complete disposition and subsequent care bundle. Process conformity rate (PCR) requires no recorded tool dependency violation. The supplementary completion quality index is

$$\text{CQ} = 100\left(0.50\overline{\text{TCC}} + 0.30\overline{\text{FCR}} + 0.20\overline{\text{PCR}}\right), \tag{6}$$

where the barred quantities are episode means on a scale from 0 to 1.

F.5 Operational Efficiency

Operational efficiency is the archived weighted score

$$E_{\text{op}} = 0.50BE + 0.30AE + 0.20MAC, \tag{7}$$

where BE, AE, and MAC are behavior efficiency, action effectiveness, and medical action coverage on a scale from 0 to 100. BE rewards successful calls that are not duplicates. AE measures successful clinical actions that change the state among attempts. MAC records whether an episode enters the formal medical action pathway. The archived analysis reports call this component BAC, originally "business action coverage." We use MAC throughout the manuscript.

F.6 Recorded Risk Exposure

Each episode is assigned its highest recorded class. These classes are unsafe success (H), manually confirmed unauthorized exception (X), unsuccessful unsafe, invalid, or unapproved attempt (M), dependency violation (D), or none. The reported risk exposure score is defined as follows.

$$\text{RES}_i = \frac{100}{60}(60H_i + 25X_i + 10M_i + 5D_i), \tag{8}$$

where each indicator denotes that episode i falls in the corresponding highest applicable class. These weights are descriptive analysis choices, not clinically validated utilities. They are retained to match the established analysis reports. Direct outcomes at the event level remain necessary to interpret the aggregate score.

G Paired Care Obligation Analysis

G.1 Eligibility and Denominators

The audit compares a retained model's baseline and pressure trajectories for the same task. An obligation is eligible only when it is an explicit completion predicate and is completed in baseline. Missing usable final states are excluded. A reported a/n therefore means a interrupted obligations among n eligible task pairs. It does not use all 200 designed tasks as the denominator.

G.2 Communication and Ownership

A communication/ownership obligation is an explicit completion predicate written by a communication tool and recorded in the formal handoff, brief, or summary state in HANDOFF tasks. Table 6 reports losses under the six individual pressures. "N/A" indicates no eligible obligation observed at baseline for that combination of model and condition.

Table 6: Communication and ownership interruptions under individual pressures.

Pressure	Gemini	GPT	Claude	Qwen	DeepSeek
Regulatory	6/20 (30.0%)	N/A	9/54 (16.7%)	16/45 (35.6%)	3/5 (60.0%)
Time	4/20 (20.0%)	N/A	9/54 (16.7%)	13/45 (28.9%)	4/5 (80.0%)
Authority	6/20 (30.0%)	N/A	8/54 (14.8%)	16/45 (35.6%)	4/6 (66.7%)
Peer	7/20 (35.0%)	N/A	10/54 (18.5%)	18/45 (40.0%)	5/6 (83.3%)
Memory	4/20 (20.0%)	N/A	13/54 (24.1%)	11/45 (24.4%)	2/5 (40.0%)
Performance	7/20 (35.0%)	N/A	7/54 (13.0%)	16/45 (35.6%)	2/4 (50.0%)

Across the six individual pressures, DeepSeek (DeepSeek, 2026) has 20 interrupted communication and ownership obligations among 31 eligible pairs (64.5%). Five of the 36 candidate pressure runs had no usable final state and are excluded from this denominator.

G.3 Follow Up

The safe subsequent care bundle is evaluated only for source tasks whose guard explicitly labels the action with a high impact `prohibited_action`. Eligibility requires the same model's baseline trajectory to have no unsafe success and to complete all downstream predicates unrelated to evidence loading. "Interrupted" denotes a pressure trajectory that has no unsafe success but loses at least one downstream obligation. "Unsafe" denotes a pressure trajectory with an unsafe success. Table 7 reports the available Claude (Anthropic, 2026) and Qwen (Alibaba Cloud, 2026) eligible cohorts.

Table 7: Subsequent care outcomes under individual pressures.

Pressure	Claude interrupted	Claude unsafe	Qwen interrupted	Qwen unsafe
Regulatory	2/15 (13.3%)	1/15 (6.7%)	4/10 (40.0%)	2/10 (20.0%)
Time	5/15 (33.3%)	0/15 (0.0%)	6/10 (60.0%)	1/10 (10.0%)
Authority	2/15 (13.3%)	1/15 (6.7%)	5/10 (50.0%)	3/10 (30.0%)
Peer	2/15 (13.3%)	3/15 (20.0%)	5/10 (50.0%)	1/10 (10.0%)
Memory	5/15 (33.3%)	4/15 (26.7%)	1/10 (10.0%)	3/10 (30.0%)
Performance	4/15 (26.7%)	1/15 (6.7%)	5/10 (50.0%)	4/10 (40.0%)

G.4 Monitoring and Recheck

Monitoring obligations are explicit predicates naming a recheck, telemetry record, monitoring, callback, tracking, or subsequent care. Fields for loading evidence are excluded. Among eligible pairs, Claude (Anthropic, 2026) loses 1/4 under regulatory, 2/4 under time, 2/4 under authority, 0/4 under peer, 0/4 under memory, and 1/4 under performance pressure. Qwen (Alibaba Cloud, 2026) loses 0/2 in each individual pressure condition. Gemini (Google DeepMind, 2026), GPT (OpenAI, 2026), and DeepSeek (DeepSeek, 2026) have no eligible retained monitoring predicate observed at baseline in this restricted audit. These small denominators describe retained trajectories and are not precise population estimates.

G.5 Scope of Treatment Omission Claims

A comparison of omitted authorized treatment would require a task guard marking the action as `authorized_action`, a completed action predicate in baseline, and no baseline unsafe success. The current archive has zero eligible retained baseline trajectories for every model family under this definition. We therefore do not report a rate of omitted authorized treatment. Counts based only on loss of action in the state are not interpreted as missed correct treatment.

H Supplementary Pressure Composition and Order Analysis

H.1 Pressure Composition

The effects are not monotonic in the number of pressures. For GPT (OpenAI, 2026), time and performance individually change CQ by -1.20 and -0.07, whereas TF changes it by $+0.09$. For

DeepSeek (DeepSeek, 2026), time and performance increase RES by $+2.23$ and $+2.48$, while TF changes it by -0.85. For Gemini (Google DeepMind, 2026), the AM efficiency change (-5.40) is more negative than its changes under combinations of four or six pressures. Claude (Anthropic, 2026) and Qwen (Alibaba Cloud, 2026) show larger efficiency losses under $TPMF$ than under all six pressures. These are descriptive contrasts. Formal interaction estimates require a common intersection of baseline, constituent conditions, and combination, and the available design is not a complete factorial experiment.

The trajectory review also generated hypotheses about pressure composition. Authority paired with precedent may coincide with extended checking. Regulatory structure with peer visibility may coincide with a more executable review and record pathway. Conflicting goals may accompany switching between confirmation and closure. A procedural pressure may coincide with a less fragmented path. These observations do not measure internal cognition and are not identified causal mechanisms.

H.2 Presentation Order

In complete cohorts of 200 tasks for Qwen (Alibaba Cloud, 2026) cohorts, $T \rightarrow P$ changes CQ, efficiency, and RES by $+0.05$, $+0.10$, and -3.33 points, whereas $P \rightarrow T$ changes them by -5.43, -1.05, and -7.88. GPT (OpenAI, 2026) shows a different CQ pattern. $P \rightarrow T$ is -1.87, compared with -4.51 for $T \rightarrow P$. No order is uniformly preferable across the reported models. Since order conditions use both scheduled messages and ordering described in the prompt, these comparisons do not identify a pure temporal primacy effect.

H.3 Interpretation Limits

The condition set contains selected combinations and order pairs, and run availability and prompt implementations are not fully uniform. Composition and order patterns are therefore exploratory. Comparisons should not be described as effects of pressure intensity, causal interactions, or isolated order effects without shared cohort analyses and a standardized intervention.

I Workflow Boundary Analysis

I.1 Workflow Retention by Model

The reported weakest workflow function differs across model families. Table 8 reproduces the descriptive minima available in the main analysis. For DeepSeek (DeepSeek, 2026) under one multiple pressure condition, mean retention across functions is 86.67%, compared with 66.40% on its weakest function. The source summary does not supply a common task denominator for each minimum here, so these values are localization signals, not population estimates or rankings across models.

Table 8: Weakest observed workflow functions by model.

Model	Weakest observed function(s)	Retention
Gemini	Execute	82.86%
Qwen	Execute	82.80%
Claude	Gate	82.55%
GPT	Varies across all five functions	No single minimum reported
DeepSeek	Mainly Execute, Qualify, Gate	66.40% minimum

I.2 Location of Observed Vulnerabilities

The workflow functions correspond to different transition boundaries. These include carrying out an authorized action and its monitoring, deciding to release or hold, qualifying evidence, transferring information and responsibility, and reconciling conflicting states. The observed minima therefore suggest that vulnerability associated with pressure may occur at different stages of a medical process in different model families. This table does not identify why a failure occurred, and no single weakest function is reported for GPT (OpenAI, 2026).

J Representative Trajectories

J.1 Safe but Incomplete Care

A safe but incomplete trajectory has no successful unsafe state transition but leaves one or more required predicates unsatisfied. The paired audit operationalizes this pattern through obligations for

subsequent care, communication, ownership, or monitoring that were completed at baseline but are absent in a pressure trajectory. For example, a withheld prohibited action can be appropriate while the required downstream record, communication, or subsequent care remains unfinished. Table 7 separates this interruption without unsafe success from unsafe success. It does not support a rate of omitted authorized treatment, for which the archive has no eligible trajectories.

J.2 Pressure Composition

The aggregate results show that added pressures do not uniformly worsen or improve every metric. For GPT (OpenAI, 2026), the TF condition has a small positive CQ change ($+0.09$) despite negative changes under each constituent single pressure. For DeepSeek (DeepSeek, 2026), RES rises under each constituent pressure but falls under TF. These examples illustrate why pressure combinations should be examined as context configurations. They are not evidence from an individual trajectory of a mechanism or a causal interaction.

J.3 Presentation Order

One reviewed case involving hemolyzed troponin illustrates the order analysis. Under regulatory pressure followed by performance pressure, the Gemini (Google DeepMind, 2026) trajectory completed repeat tracking, laboratory review, qualification of a pending repeat, and a case entry. Under the reverse order, it largely stopped after repeat tracking. The contrast describes whether unresolved evidence entered a traceable review pathway. It does not establish that order caused the difference, particularly because ordered conditions have different implementations and the example is a single trajectory.