BehaviorSFT: Behavioral Token Conditioning for Health Agents Across the Proactivity Spectrum

Yubin Kim[1], Zhiyuan Hu[1,2], Hyewon Jeong[1], Eugene W Park[1], Shuyue Stella Li[3],
Chanwoo Park[1], Shiyun Xiong[2], MingYu Lu[3], Hyeonhoon Lee[4], Xin Liu[5],
Daniel McDuff[5], Cynthia Breazeal[1], Samir Tulebaev[6], Hae Won Park[1]
[1]Massachusetts Institute of Technology, [2]National University of Singapore
[3]University of Washington, [4]Seoul National University Hospital
[5]Google Research, [6]Mass General Brigham
Correspondence: ybkim95@mit.edu

Abstract

Large Language Models (LLMs) as agents require careful behavioral adaptation. While adept at reactive tasks (e.g., medical reasoning), LLMs often struggle with proactive engagement, like unprompted identification of critical missing information or risks. We introduce **BEHAVIORBENCH**, a comprehensive dataset to evaluate agent behaviors across a clinical assistance spectrum. To rigorously test the current models, we also introduce **BEHAVIORBENCH-HARD**, a challenging subset where the performance of state-of-the-art models drops significantly, revealing weaknesses. To address these challenges, we propose **BEHAVIORSFT**, a novel training strategy using behavioral tokens to explicitly condition LLMs for dynamic behavioral selection which boosts performance on both benchmarks. Crucially, a blind clinician evaluation confirmed that our trained agents exhibit more realistic clinical behavior, striking a superior balance between helpful proactivity and necessary restraint versus standard fine-tuning or explicitly instructed agents.[1]

1 Introduction

As Large Language Models (LLMs) transition from experimental systems to deployed *agents* in clinical environments (Heydari et al., 2025; Khasentino et al., 2025), a critical question emerges: "*when* and *how* should these systems act *reactively* or *proactively* (Fauscette, 2024)?". Unlike general-purpose AI agents, health agents can operate in high-stakes environments where both action and inaction carry significant consequences (Kim, 2025; Kim et al., 2025a,c). We define *reactive* behaviors as those where the agent responds only to explicit queries with precisely the information requested, while *proactive* behaviors involve volunteering additional information, raising concerns, or suggesting actions beyond what was directly solicited. Importantly, proactivity in clinical contexts extends beyond merely asking clarifying questions, a common, but limited, focus in existing NLP research (Li et al., 2024; Hu et al., 2024). While question-asking represents one dimension of proactivity, our work encompasses a broader spectrum; including unsolicited intervention, critical evaluation, and recommendation. These behaviors align closely with the "Appraisal" phase of Evidence-Based Medicine (EBM) (Denby, 2008), where clinicians actively assess available information, identify information gaps, and determine appropriate next steps. An agent that remains strictly reactive may fail to raise an alert when problems are observed with critical lab values or medication contraindications (Walter Costa et al., 2021; Wright et al., 2018), potentially compromising patient safety (McCoy et al., 2014). In contrast, an excessively proactive system that frequently interrupts with unsolicited recommendations risks contributing to alert fatigue, interruption of workflow, and potential rejection by health professionals (Sutton et al., 2020). This trade-off between reactive and proactive behaviors forms the core challenge addressed in this paper. The appropriate balance between these modalities varies dramatically based on clinical context, urgency, risk levels, and the specific health roles being augmented, demanding adaptive behavior policy rather than a fixed mode, especially as systems achieve higher levels of autonomy (Figure 4).

To systematically discuss how an agent's reactive and proactive stance should adapt with its increasing capabilities, we adapt the SAE Levels of Driving Automation (SAE, 2021) into a six-level taxonomy for health agent autonomy. This framework detailed in Table 10 helps to illustrate a key principle: as an agent ascends these autonomy levels, its capacity and responsibility to engage in sophisticated proactive behaviors, rather than merely

[1]Project Page: https://behavior-adaptation.github.io/

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 9789–9817
November 4-9, 2025

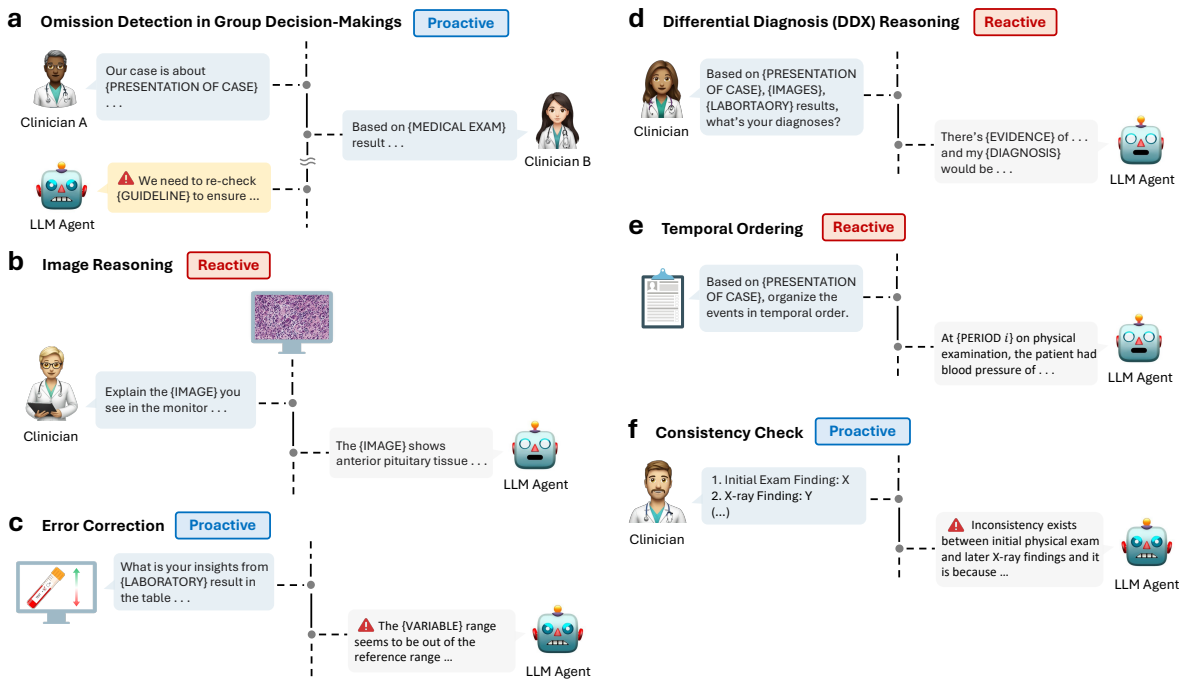

Figure 1: **Six representative tasks from BEHAVIORBENCH, showcasing the spectrum of agent behaviors in clinical settings.** The figure illustrates (a-c, f) proactive tasks where the LLM agent identifies issues or offers insights without direct prompting, and (b, d, e) reactive tasks responding to explicit clinician queries.

reactive ones, become increasingly critical.

The autonomy level taxonomy highlights that effective health AI, particularly for achieving **Level 3 (Conditional Proactive Assistance) and above**, must move beyond simple reactive responses (Levels 1-2). As AI autonomy increases, the nature of clinician responsibility evolves, shifting from direct task execution to supervision, validation of AI-driven insights, and management of exceptions. Our work, therefore, focuses on enabling AI agents to learn and exhibit the adapted spectrum of reactive and proactive behaviors crucial for safe and effective operation at these higher levels of conditional and collaborative automation. BEHAVIORBENCH{-HARD} are designed to evaluate these capabilities across this spectrum, and BEHAVIORSFT aims to train agents to achieve this behavioral adaptability, particularly for robust performance at Levels 2 and 3, with an eye towards future capabilities at Level 4.

Effectively adapting *which* of these behaviors is appropriate, and *when*, it is essential for health AI systems that can safely operate at increasing levels of autonomy. In this work, we ask "*what proactivity means for health AI and how we build systems that are appropriately behaving?*". To this end, we propose a novel six-level taxonomy for health AI autonomy that maps progression from human-controlled to autonomous operation. We trace the evolution from early reactive systems (Tu et al., 2024; Han et al., 2023) to more recent developments like MediQ (Li et al., 2024), AIME (McDuff et al., 2025; Tu et al., 2024) and Tiered Agentic Oversight (Kim et al., 2025b), which incorporate proactive elements while demonstrating the critical interplay between proactivity and urgency. Our benchmark was curated from real medical cases sourced from New England Journal of Medicine (NEJM) clinical case reports (Brinkmann et al., 2024). We employed LLMs (`Gemini-2.5 Flash` and `Gemini-2.5 Pro`) to meticulously ground these cases in their factual details and then reformat them into multi-turn, multi-clinician-patient conversational scenarios, integrating multi-modal inputs such as text, images, and tabular data. Indeed, we propose this LLM-assisted methodology for converting existing static clinical datasets into rich, reactive-proactive benchmark scenarios as a key contribution of our work. Additionally, we present a novel training methodology, BEHAVIORSFT, which employs explicit behavioral tokens to condition LLM responses along the reactive-proactive spectrum. Our approach demonstrates significant improvements, achieving up to 97.3%

overall Macro F1 on BehaviorBench (compared to 96.7% for general SFT) with particularly notable gains in proactive tasks (from 95.0% to 96.5%). The primary contributions are:

1. We introduce BEHAVIORBENCH{-HARD}, evaluation datasets that assesses LLM capabilities across both reactive and proactive tasks in health contexts.

2. We provide detailed analysis of recent LLMs' performance on our benchmarks, revealing significant variability in contextual awareness and appropriate behavioral adaptation.

3. We propose BEHAVIORSFT, a new fine-tuning strategy that leverages behavioral tokens to guide LLMs in dynamically adapting their responses along the reactive-proactive spectrum tasks.

2 BEHAVIORBENCH

We introduce BEHAVIORBENCH, a novel dataset specifically designed to assess agent capabilities across the reactive-proactive tasks. Derived from real clinical cases, BEHAVIORBENCH comprises of 6,876 real-world clinical case scenarios from which we derived a total of 142,496 tasks distributed across the 13 distinct task categories. This framework provides a more granular analysis of an agent's ability to discern context and modulate its behavior accordingly, moving beyond standard metrics, such as accuracy, that are solely based on reactive responses. Detailed dataset statistics can be found in the Appendix C.

To ensure that the generated tasks effectively probe clinical reasoning, we construct the dataset in a two-step process. First, we carefully prompt the LLM (see Appendix F) generating the tasks to use detailed summary from real-world clinical cases, including patient history, diagnostics, conversation snippets, and final diagnoses. This ensures that the questions, answers, and rationales reflect genuine clinical context instead of relying on pseudolabels generated without any realistic groundings. All draft tasks then underwent several back-and-forth revision cycles with two physicians, who reviewed any hallucinations and confirmed each scenario's practical plausibility for N=10 cases. Then, to evaluate the agent's proactive capabilities, we augment the base scenarios by intentionally introducing subtle challenges, such as hypothetical scenarios with probable clinical errors, conflicting data points (e.g. modifying numerical values slightly between reports, or presenting exam findings seemingly at odds with imaging), and omitted information expected by clinical standards. The resulting reactive-proactive tasks are as follows:

Reactive Tasks evaluates whether the agents can handle information when requested directly.

1. `fact_retrieval`: Finds specific facts mentioned in the text (e.g., "What was the patient's initial temperature?").

2. `timeline_sequence`: Puts events in order using clear time references (e.g., tracing how lung exam findings changed between the initial presentation and Turn N, based on provided descriptions from those time points).

3. `ddx_reasoning`: Explains the reasoning for a possible diagnosis using only the evidence given (e.g., identifying findings prior to Turn M, such as specific X-ray descriptions and sputum results, that suggested bronchopneumonia over simple lobar pneumonia).

4. `treatment_decision`: Connects a doctor's thinking or action to the stated reason or data supporting it (e.g., evaluating a specific diagnostic leaning mentioned in Turn K based only on the evidence explicitly available at that time, like sputum results).

Balanced Tasks are initiated by specific, provided information but demand a more significant cognitive step involving deeper thinking, such as multi-step inference, synthesis of multiple data points, or evaluating the impact of new information on existing understanding.

1. `reasoning_differential_evolution`: Compares the patient's situation at two different times and explains how the doctor's assessment should change because of new information (e.g., asking how the list of possible diagnoses should shift from Timepoint A to Timepoint B considering newly available sputum culture results and vital signs).

2. `integrity_missing_turn_inference`: Figures out what was likely said in a missing part of a conversation based on what came before and after (e.g., "Turn N orders a test, Turn $N + M$ discusses the result. What

Table 1: **Comparison of Public Medical Benchmarks.** Modality codes: *t*=text, *i*=image, *b*=tabular/structured data. ✓ indicates that the benchmark natively supports the evaluation dimension; ✗ indicates it does not.

Benchmark	Size	Modality	Behavior Evaluation	Sequential Eval.	Dialogue Interaction	Multiple Roles
MedQA (Jin et al., 2021)	1,273	*t*	✗	✗	✗	✗
MedMCQA (Pal et al., 2022)	6,100	*t*	✗	✗	✗	✗
MultiMedQA (Singhal et al., 2023)	13,115	*t*	✗	✗	✗	✗
MediQ (Li et al., 2024)	1,273	*t*	✗	✓	✓	✓
MediQ-AskDocs (Li et al., 2025a)	17,000	*t*	✗	✓	✓	✓
ClinicBench (Chen et al., 2024)	11,000	*t*	✗	✗	✗	✗
MedChain (Liu et al., 2024)	12,163	*t+i*	✗	✓	✓	✓
MedAgentBench (Jiang et al., 2025)	300	*t+b*	✗	✓	✓	✓
HealthBench (Arora et al., 2025)	5,000	*t*	✓	✗	✓	✗
BEHAVIORBENCH (Ours)	142,496	***t+i+b***	✓	✓	✓	✓

likely happened in Turn $N + K$, where $0 < K < M$?").

Proactive Tasks require the LLM to use higher-level thinking, and evaluation skills.

1. `predictive_next_action`: Forecasts the most appropriate subsequent clinical action by integrating the evolving patient case, current symptoms, medical history, and available diagnostic results.

2. `explicit_error_correction`: Identifies and rectifies explicitly stated errors in clinical narratives or proposed actions, providing justifications based on medical knowledge and case specifics (e.g., correcting drug suitability given a patient's allergy).

3. `omission_detection`: Identifies significant omissions in the provided clinical information or documented actions, such as overlooked diagnostic tests or unaddressed critical symptoms that could impact patient care.

4. `standard_of_care`: Assesses whether documented clinical management, including diagnostic procedures and interventions, adheres to established medical guidelines and accepted best practices, often requiring external knowledge.

5. `interpretation_conflict`: Discerns and reconciles nuanced or potentially conflicting interpretations of clinical findings from different sources (e.g., contrasting physical exam notes with radiology findings), articulating their clinical significance.

6. `data_conflict_resolution`: Identifies direct contradictions or inconsistencies between pieces of factual clinical data presented within a case (e.g., conflicting lab values over time) and proposes logical explanations.

7. `consistency_check`: Evaluates the overall logical and clinical coherence of a case narrative or specific information, identifying elements that are incongruous or implausible (e.g., assessing if a patient's reported progression aligns with a given diagnosis).

BEHAVIORBENCH-HARD To further enhance the difficulty of our benchmark, we curated BEHAVIORBENCH-HARD, a subset of 297 challenging cases where multiple state-of-the-art models consistently fail. We selected cases by evaluating three reasoning-specialized LLMs (o3, Gemini-2.5-Pro, DeepSeek-R1) on the full test set with three random seeds each, choosing instances where $\geq$2 models failed across multiple runs. The resulting subset maintains task distribution balance (43% proactive, 39% reactive, 18% balanced).

3 BehaviorSFT: Behavior Adaptation Training

To operationalize the concept of behavioral adaptation within health LLM agents, we propose a targeted training strategy, Behavior-Conditioned Supervised Fine-Tuning (BehaviorSFT). This approach leverages our specialized `BehaviorBench` dataset (Section 2) to explicitly teach LLMs to modulate their responses along the reactive-proactive spectrum based on inferred clinical context. This contrasts with standard SFT approaches, which typically optimize for task completion without explicit mechanisms to control the agent's level of initiative or caution, risking either unsafe passivity or disruptive over-intervention.

3.1 Behavior Tokens

Rationale for Prefix Tokens: We employ prefix behavior tokens (e.g., <reactive>, <proactive>) for several reasons. Placing the token at the beginning of the target sequence allows it to act as a direct control signal, conditioning the entire generation process on the desired behavioral mode from the outset. This explicitly trains the model to adopt the appropriate style, tone, and level of initiative as it generates the response. While one could consider predicting the token after some internal reasoning chain, our approach integrates this reasoning implicitly, i.e., the model learns to predict the correct initial token based on its understanding of the input context (x), as described in our Contextual Behavior Assessment capability (Section 3.3). This provides an end-to-end mechanism for context-aware, behaviorally adapted generation. The key to our approach is the special behavior tokens paired with the target response during training.

- `<reactive>`: Signals the generation of a direct, concise response strictly adhering to the explicit query, avoiding unsolicited information or inferences.
- `<proactive>`: Signals a response that may include identifying implicit issues, volunteering relevant context or warnings, suggesting next steps, or applying external knowledge (e.g., standards of care) beyond the literal query.

These tokens act as control signals, learned by the model and conditioning the subsequent generation process. Alternative approaches exist, such as training a separate classifier to select the mode and then routing the input to specialized reactive or proactive models, or using inference-time techniques like thresholding logits associated with the behavior tokens for finer control. However, our BehaviorSFT approach offers a simpler, unified training process within a single model. Future work could explore hybrid methods or compare the efficacy of these different control paradigms.

3.2 Training Data

`BehaviorBench` serves as the crucial training ground for BehaviorSFT. Each instance within the benchmark's training split is meticulously annotated with the desired target behavior token based on the task's nature and the underlying clinical scenario's demands:

1. **Reactive Annotation (`<reactive>`):** Applied to tasks demanding factual recall, direct sequencing, or simple reasoning strictly from provided data (e.g., `fact_retrieval`, `timeline_sequence`).
2. **Proactive Annotation (`<proactive>`):** Applied to tasks necessitating critical assessment, error/omission detection, consistency checking, or prediction based on clinical standards (e.g., `consistency_check`, `standard_of_care`, `predictive_next_action`).
3. **Contextual Annotation for Balanced Tasks:** Instances from balanced tasks (e.g., `reasoning_differential_evolution`) are annotated based on whether the specific context warrants simple reporting (`<reactive>`) or highlighting significant changes/implications (`<proactive>`).

Each annotated instance is then structured for auto-regressive SFT, pairing the input context/query with a target sequence beginning with the assigned behavior token, followed by an ideal response exemplifying that behavior.

Example 1 (Reactive Task):

```
Input: Context: [Note excerpt: Vitals
    stable.]
Query: Latest vitals?
Target: <reactive> BP 120/80, HR 75,
    Temp 37.0C, RR 16.
```

Example 2 (Proactive Task):

```
Input: Context: [Chart: Rx Drug A.
    Allergy list: Drug A.]
Query: Confirm med list okay?
Target: <proactive> Warning: Drug A
    prescribed but patient is allergic.
    Review immediately.
```

This structured data format explicitly teaches the model the association between clinical scenarios, appropriate behavioral modes (reactive/proactive), and corresponding linguistic outputs.

3.3 Training Procedure: BehaviorSFT

Starting with a pre-trained foundation LLM, we perform SFT using the behavior-annotated `BehaviorBench` training data. The objective is the standard causal language modeling loss, minimizing the negative log-likelihood of the target sequence $y = (y_1, ..., y_T)$, where $y_1 \in$

{<reactive>, <proactive>}:

$$\mathcal{L}_{BehaviorSFT} = -\sum_{i=1}^{T} \log P(y_i | y_{<i}, x; \theta) \quad (1)$$

Here, x is the input context/query, $y_{<i}$ are the preceding target tokens, and θ represents the model parameters.

Through this process, the model learns the crucial, intertwined capabilities:

1. **Contextual Behavior Assessment:** Implicitly analyzing the input x to determine the likelihood that a proactive or reactive stance is warranted, influencing the prediction of the initial token y_1.

2. **Behavior-Conditioned Generation:** Generating subsequent tokens $y_{2:T}$ in a manner consistent with the generated or given behavior token y_1, adopting the appropriate style, tone, and level of detail or intervention.

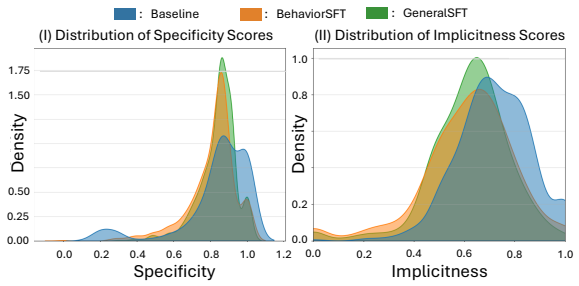

Figure 2: **Density distributions of (I) Specificity and (II) Implicitness scores for Baseline, BehaviorSFT, and GeneralSFT agent outputs.** (I) Specificity: Both fine-tuned models (BehaviorSFT and GeneralSFT) markedly improve output specificity over the Baseline, with distributions concentrated at high scores (∼0.9). (II) Implicitness: Distinct implicitness profiles emerge: GeneralSFT is the most explicit (lowest scores, ∼0.6-0.7), the Baseline is the most implicit (highest scores, ∼0.7-0.9), while BehaviorSFT exhibits a moderate, intermediate level of implicitness (∼0.7-0.8).

4 Experiments and Results

4.1 Setup

All experiments use BEHAVIORBENCH with fixed `6776/110/977` train–val–test split. We fine–tune both backbones; `Qwen-2.5-7B-Instruct` (Team, 2024) and `Meta-Llama-3.1-8B-Instruct` (Meta AI, 2024). Implementation details can be found in Appendix G.

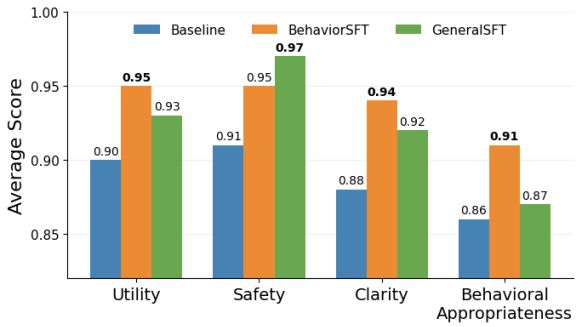

Figure 3: **G-Eval with `gpt-4o-mini` as evaluator of `Qwen-2.5-7B-Ins` responses across four key metrics.** We compare the average scores for the Baseline model, our proposed BehaviorSFT, and GeneralSFT. BehaviorSFT consistently outperforms the Baseline across all metrics and demonstrates competitive or superior performance compared to GeneralSFT.

4.2 Main Results

From Reactive to Proactive capabilities in clinical LLMs involve processing and responding directly to explicitly provided information. Reactivity encompasses *fact retrieval*, *information summarization*, ordering events via *direct sequencing*, following *simple execution* instructions, and performing *basic reasoning from explicit data*, these tasks test the LLM's ability to understand and manipulate information as presented, without significant inference or applying external knowledge. The Proactive-Reactive Scale of 0.0-0.4 typically reflects these functions.

Conversely, require the LLM to transcend literal interpretation, demonstrating deeper reasoning, anticipation, and critical assessment. Key aspects include *inference and implication* (identifying unstated assumptions or missing information), *anticipation and prediction* (foreseeing next steps or complications), *consistency and conflict detection* (finding discrepancies between data points), *error recognition and correction*, *applying external knowledge* like standards of care, and *synthesis and complex interpretation* from multiple sources. These tasks simulate higher-order clinical thinking. The Proactive-Reactive Scale of 0.6-1.0 aligns with these skills, while 0.4-0.6 represents a balance.

Empirical Results Overview. Table 2 reports Macro F1 scores across the three task categories on BEHAVIORBENCH. Relative to both the majority-voting Ensemble baseline and standard supervised fine-tuning (Gen. SFT), BehaviorSFT matches or exceeds performance on the Reactive and Balanced sets, and yields a clear advantage on

Table 2: **Performance on BEHAVIORBENCH.** We report Macro F1-scores (%) across three task categories. Best result per task is highlighted in **bold**.The **Ensemble** column reports baseline performance by majority voting across three commercial closed-source models (`Gemini-2.5-pro`, `OpenAI-o1`, `DeepSeek-R1`). 'ZS' = Zero-Shot, 'FS (k=3)' = Few-Shot (3 examples), 'CoT' = Chain-of-Thought, 'Explicit Instr.' = ZS with explicit reactive/proactive instruction, 'Gen. SFT' = Standard Supervised Fine-Tuning (SFT), 'BehaviorSFT' = Our proposed fine-tuning method.

Category	Task	Ensemble			Qwen2.5-7B-Ins		Llama3.1-8B-Ins	
		ZS	FS (k=3)	ZS + Explicit Instr.	Gen. SFT	BehaviorSFT	Gen. SFT	BehaviorSFT
Reactive	fact_retrieval	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	timeline_sequence	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	ddx_reasoning	96.2	96.6	96.6	96.1	96.1	94.2	92.7
	treatment_decision	94.8	95.3	95.3	100.0	98.4	98.4	98.7
	Avg.	98.2	98.2	98.2	**98.6**	**98.6**	97.8	97.2
Balanced	reasoning_diff_evolution	98.6	98.6	98.6	100.0	100.0	100.0	100.0
	integrity_missing_turn	100.0	100.0	100.0	100.0	100.0	96.4	100.0
	Avg.	97.2	97.6	97.6	**100.0**	99.2	98.5	**100.0**
Proactive	consistency_check	94.3	100.0	94.3	100.0	100.0	100.0	100.0
	data_conflict_resolution	97.2	97.2	97.2	99.3	98.6	99.2	98.6
	interpretation_conflict	98.5	96.5	96.5	96.6	96.6	98.5	98.6
	standard_of_care	93.4	95.3	93.7	94.8	93.3	91.5	88.4
	omission_detection	89.5	92.4	89.3	88.5	95.1	90.0	93.2
	explicit_error_correction	96.3	97.5	96.4	98.3	99.2	98.4	97.2
	predictive_next_action	82.5	83.0	82.3	84.8	91.7	77.0	83.4
	Avg.	94.3	95.1	94.0	95.0	**96.5**	94.2	94.7
Avg.		95.4	96.0	95.3	96.7	**97.3**	95.8	96.1

Table 3: Macro F1-scores of prompting methods on behavior classification. Method abbreviations: **BT** = Behavior token, **BC** = Behavior chain-of-thought, **OC** = Option CoT, **OP** = Option. Class abbreviations: **Five-class** (BA = balanced; H_PR = highly_proactive; H_RE = highly_reactive; P_PR = primarily_proactive; P_RE = primarily_reactive), **Binary** (PR = proactive; N_PR = non-proactive), **Three-class** (BA = balanced; PR = proactive; RE = reactive).

	Five-class					Binary		Three-class		
	BA	H_PR	H_RE	P_PR	P_RE	PR	N_PR	BA	PR	RE
BT-OC-OP	42.62	89.47	4.76	19.19	68.72	82.14	92.10	53.41	92.10	73.68
BT-OP	37.06	87.77	13.79	25.28	66.40	82.76	92.19	46.92	92.19	66.42
BT-BC-OC-OP	58.24	87.84	19.05	11.82	71.75	83.48	92.90	51.67	92.90	72.09
BT-BC-OP	54.74	88.89	17.39	11.00	73.68	82.97	92.58	51.76	92.58	69.57
BC-BT-OC-OP	57.06	87.73	14.81	7.07	74.89	82.59	92.23	45.00	92.32	69.96

the most demanding Proactive tasks (Qwen: 96.5% vs. 95.0%; Llama: 94.7% vs. 94.2%). The benefits of this behavior-aligned strategy are magnified on BEHAVIORBENCH-HARD, our challenging subset detailed in Table 4. On these adversarial cases, all models experienced a significant performance drop, but BehaviorSFT demonstrated superior robustness, achieving the highest overall F1-score of 73.6% compared to 71.3% for GeneralSFT. This confirms that our approach not only improves performance on average but also builds more resilient agents for complex clinical reasoning. Detailed performance for larger baselines is in Appendix E.

Enhanced User-Centric Qualities with G-Evaluation Our evaluation using G-Eval (Liu et al., 2023), a methodology leveraging large models for human-aligned assessment, reveals significant qualitative improvements with BehaviorSFT. As depicted in Figure 3, BehaviorSFT consistently outperforms the Baseline across all four key metrics: Utility, Safety, Clarity, and Behavioral Appropriateness. Notably, BehaviorSFT achieves the highest scores in Utility (0.95 vs. 0.93 for GeneralSFT and 0.90 for Baseline), Clarity (0.94 vs. 0.92 for GeneralSFT and 0.88 for Baseline), and Behavioral Appropriateness (0.91 vs. 0.87 for Gen-

Table 4: F1-scores on BEHAVIORBENCH-HARD and comparsion with BEHAVIORBENCH. The Ensemble column reports baseline performance by majority voting across three closed-source models (Gemini-2.5-pro, OpenAI-o1, and DeepSeek-R1). All models show significant degradation, with BehaviorSFT demonstrating superior robustness across most of task categories. Here, the GeneralSFT and BehaviorSFT is based on Qwen2.5-7B-Ins model.

Model	BEHAVIORBENCH				BEHAVIORBENCH-HARD			
	Overall	Reactive	Balanced	Proactive	Overall	Reactive	Balanced	Proactive
Ensemble (ZS)	95.4	98.2	97.2	94.3	68.9	76.5	71.2	62.8
GeneralSFT	96.7	**98.6**	**100.0**	95.0	71.3	79.2	73.8	65.1
BehaviorSFT	**97.3**	**98.6**	99.2	**96.5**	**73.6**	**81.3**	**76.2**	**68.4**

eralSFT and 0.86 for Baseline). While GeneralSFT scores marginally higher in Safety (0.97 vs. 0.95 for BehaviorSFT), BehaviorSFT still demonstrates a strong safety profile. These results underscore BehaviorSFT's capability to not only perform tasks effectively but also to align more closely with user expectations in terms of usefulness, understandability, and appropriate interaction, suggesting a more refined and user-centric agent behavior.

Optimizing Output Specificity while Balancing Implicitness Figure 2 illustrates the impact of our fine-tuning approaches on the nuanced characteristics of agent responses, specifically their specificity and implicitness. Both fine-tuned models, BehaviorSFT and GeneralSFT, markedly enhance output specificity compared to the Baseline, with distributions concentrating at high specificity scores (around 0.9). This indicates that both methods generate more detailed and precise information. However, a key distinction emerges in their implicitness profiles. GeneralSFT tends towards more explicit communication, reflected in lower implicitness scores (approximately 0.6-0.7). In contrast, the Baseline model is the most implicit (scores around 0.7-0.9). BehaviorSFT carves out an intermediate and potentially more versatile profile, achieving a moderate level of implicitness (scores approximately 0.7-0.8). This suggests that BehaviorSFT can deliver highly specific information without resorting to excessive explicitness, potentially mirroring more natural human communication patterns and aligning with the idea that effective agents must navigate implicit evaluation criteria (Wadhwa et al., 2025).

4.3 Ablation on prompting variants for Behavior Pattern Analysis

Table 3 evaluates five prompting recipes obtained by incrementally adding *Behavior Chain-of-Thought* (BC) and *Option reasoning* (OC/OP) on top of the *Behavior Token* (BT) baseline. The full recipe *BT–BC–OC–OP* achieves the best or second-best Macro F1 in 11 of the 13 columns (e.g., *Five-class BA* 58.2 and *Binary PR* 83.5), showing that BC and OC/OP provide complementary gains. Dropping OC/OP (*BT–BC–OP*) or BC (*BT–OP*) consistently lowers scores, while reversing the BC placement (*BC–BT–OC–OP*) yields a smaller benefit, indicating that BC is most effective when appended after the BT prompt. Overall, combining both reasoning cues delivers the most robust behaviour classification across all label granularities.

5 Clinician-in-the-loop Evaluation

To rigorously evaluate our BehaviorSFT agent and validate the proposed dataset, we conducted a user study involving board-certified medical professionals. This study was designed to assess the clinical utility of BEHAVIORBENCH and to compare the performance of LLM agents exhibiting distinct behavioral characteristics. A detailed setup (participants, study procedure, and the annotation interface) of our user study can be found at H.

5.1 Annotation Results

This section presents the quantitative and qualitative findings from the clinician-in-the-loop evaluation study. All reported inter-annotator agreement scores were calculated among the three participating physicians.

5.1.1 Phase 1: BEHAVIORBENCH Validation

Clinicians evaluated a total of 60 unique tasks from the BEHAVIORBENCH.

MCQ Accuracy and Task Plausibility The physician annotators demonstrated a high level of accuracy in answering the multiple-choice questions, achieving an overall correctness of **83.3%**. This proficiency underscores their expert understanding of the clinical scenarios presented within the dataset.

The clinical plausibility of the tasks was a key validation metric. As shown in Figure 18, a substantial majority of tasks (**80.0%**) were rated as clinically plausible ("Yes"). No tasks (0.0%) were rated as definitively "No" for plausibility, while 20.0% were marked as "Unsure," suggesting areas where task framing or context might warrant further refinement or clarification for some annotators.

Annotator Confidence Levels Annotator confidence in their selected MCQ answers was recorded on a three-point scale. The distribution, illustrated in Figure 18, reveals that physicians were predominantly "High" in their confidence (55.0%). "Moderate" confidence was reported for 36.7% of answers, while "Low" confidence was expressed for only 8.3% of answers. This general trend towards higher confidence aligns with the observed accuracy.

Inter-Annotator Agreement for Dataset Validation To ensure the reliability of the dataset validation process, inter-annotator agreement was quantified using the Intraclass Correlation Coefficient (ICC3) for continuous ratings.

The task proactivity/reactivity slider ratings (0.0-1.0 scale) demonstrated *good* reliability with an ICC3 of 0.61. This robust agreement scores indicate that the physicians interpreted and applied the validation criteria consistently.

5.1.2 Phase 2: Comparative Agent Behavior Evaluation Results

Physicians evaluated agent responses across N=24 unique clinical tasks. The anonymized agents evaluated were BehaviorSFT, General SFT, and ZS + Explicit Instr.

Agent Response Ranking and Proactivity/Reactivity Appropriateness The primary evaluation involved ranking the three agents. Agent A (BEHAVIORBENCH) received the most favorable rankings, achieving the lowest (best) mean rank of 1.80 (Figure 19). In terms of the appropriateness of proactivity/reactivity, Agent C (ZS + Explicit Instr.) scored highest with a mean Likert score of 4.20 out of 5 (Figure 19). Agent B (General SFT) had a mean rank of 2.08 and a mean Likert score of 4.08.

6 Limitations and Future Works

Data & Task Scope. BEHAVIORBENCH aggregates 6,876 English clinical vignettes (142K task instances) from NEJM. This corpus reflects an internal-medicine bias and omits modalities such as radiology reads, nursing shift notes, tele-health transcripts, and non-English documentation. The future tasks include expanding the benchmark to multilingual EHR snippets and image-grounded prompts, and we are adding tasks for dermatology, psychiatry, and longitudinal trend summarisation to test whether proactive cues generalise beyond text-only, single-visit encounters.

Behavior Modeling. Our BEHAVIORSFT controller currently toggles generation with binary `<reactive>`, `<proactive>` tokens. Although effective for coarse behavior shifts, these cannot express nuances such as anticipatory clarification versus high-urgency escalation, and it occasionally over-fires, creating alert fatigue. We are experimenting with a hierarchical token inventory (e.g. `<flag_safety>`, `<escalate_critical>`) learnt from multi-label supervision, and with behaviour-weighted RLHF that continuously trades helpfulness against cognitive load.

Evaluation & Deployment Readiness. The clinician study in Section 5 involved three medical doctors and a number of cases sufficient for validation but under-powered for robust error stratification or workflow integration. Future work should recruit multi-institution cohorts (20+ clinicians, 1,000+ cases) and embeds the agent inside a simulated EHR sandbox to observe interrupt patterns, hand-off continuity, and long-horizon reasoning across multi-day episodes.

7 Conclusion

In this paper, we introduce **BEHAVIORBENCH**, a benchmark validated by clinicians, which reveals key proactivity gaps in current LLMs. To bridge these gaps, we proposed **BehaviorSFT**, a novel fine-tuning strategy using explicit behavioral tokens. Our method achieved state-of-the-art performance on BEHAVIORBENCH, with a Macro F1 of up to 97.3%. Crucially, the strength of BehaviorSFT was highlighted on BEHAVIORBENCH-HARD. While all models experienced a performance drop, BehaviorSFT demonstrated superior resilience, achieving an F1-score of 73.6%, outperforming both GeneralSFT and the ensemble baseline. In a blind user study, clinicians ranked the our trained agent as the most effective (best mean rank 1.80). Combined with best G-Eval scores for Utility and Behavioral Appropriateness, our findings show that BehaviorSFT produces more reliable, clinically nuanced, and expert-preferred agents.

References

2021. Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. Surface Vehicle Recommended Practice.

Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, and 1 others. 2025. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*.

Rory Brinkmann, Eric Rosenberg, David N Louis, and Scott H Podolsky. 2024. Building a community of medical learning—a century of case records of the massachusetts general hospital in the journal.

Canyu Chen, Jian Yu, Shan Chen, Che Liu, Zhongwei Wan, Danielle Bitterman, Fei Wang, and Kai Shu. 2024. Clinicalbench: Can llms beat traditional ml models in clinical prediction? *arXiv preprint arXiv:2411.06469*.

Maximillian Chen, Xiao Yu, Weiyan Shi, Urvi Awasthi, and Zhou Yu. 2023. Controllable mixed-initiative dialogue generation through prompting. *arXiv preprint arXiv:2305.04147*.

Eleni Chiou, Francesco Giganti, Shonit Punwani, Iasonas Kokkinos, and Eleftheria Panagiotaki. 2020. Harnessing uncertainty in domain adaptation for mri prostate lesion segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*, pages 510–520. Springer.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*.

Donald J Denby. 2008. *Evidence based medicine*. Xulon Press.

Michael Fauscette. 2024. Agentic ai vs. llms: Understanding the shift from reactive to proactive ai. https://www.arionresearch.com/blog/. Arion Research Blog.

Alexander Fixler, Blake Oliaro, Marshall Frieden, Christopher Girardo, Fiona A Winterbottom, Lisa B Fort, and Jason Hill. 2023. Alert to action: implementing artificial intelligence–driven clinical decision support tools for sepsis. *Ochsner Journal*, 23(3):222–231.

Stephen H Friend, Geoffrey S Ginsburg, and Rosalind W Picard. 2023. Wearable digital health technology.

Illin Gani, Ian Litchfield, David Shukla, Gayathri Delanerolle, Neil Cockburn, and Anna Pathmanathan. 2025. Understanding "alert fatigue" in primary care: Qualitative systematic review of general practitioners attitudes and experiences of clinical alerts, prompts, and reminders. *Journal of Medical Internet Research*, 27:e62763.

Senay A Gebreab, Khaled Salah, Raja Jayaraman, Muhammad Habib ur Rehman, and Samer Ellaham. 2024. Llm-based framework for administrative task automation in healthcare. In *2024 12th International Symposium on Digital Forensics and Security (ISDFS)*, pages 1–7. IEEE.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, and 1 others. 2025. Towards an ai co-scientist. *arXiv preprint arXiv:2502.18864*.

Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, Aditya Krishna Menon, Sanjiv Kumar, and Vaishnavh Nagarajan. 2023. Think before you speak: Training language models with pause tokens. *arXiv preprint arXiv:2310.02226*.

Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. 2023. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*.

A Ali Heydari, Ken Gu, Vidya Srinivas, Hong Yu, Zhihan Zhang, Yuwei Zhang, Akshay Paruchuri, Qian He, Hamid Palangi, Nova Hammerquist, and 1 others. 2025. The anatomy of a personal health agent. *arXiv preprint arXiv:2508.20148*.

Zhiyuan Hu, Chumin Liu, Xidong Feng, Yilun Zhao, See-Kiong Ng, Anh Tuan Luu, Junxian He, Pang Wei Koh, and Bryan Hooi. 2024. Uncertainty of thoughts: Uncertainty-aware planning enhances information seeking in large language models. *arXiv preprint arXiv:2402.03271*.

Mustafa I Hussain, Tera L Reynolds, and Kai Zheng. 2019. Medication safety alert fatigue may be reduced via interaction design and clinical role tailoring: a systematic review. *Journal of the American Medical Informatics Association*, 26(10):1141–1149.

Yixing Jiang, Kameron C Black, Gloria Geng, Danny Park, Andrew Y Ng, and Jonathan H Chen. 2025. Medagentbench: Dataset for benchmarking llms as agents in medical applications. *arXiv preprint arXiv:2501.14654*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.

Mohamed Khalifa and Mona Albadawy. 2024. Artificial intelligence for clinical prediction: exploring key domains and essential functions. *Computer Methods and Programs in Biomedicine Update*, page 100148.

Justin Khasentino, Anastasiya Belyaeva, Xin Liu, Zhun Yang, Nicholas A Furlotte, Chace Lee, Erik Schenck, Yojan Patel, Jian Cui, Logan Douglas Schneider, and 1 others. 2025. A personal health large language model for sleep and fitness coaching. *Nature Medicine*, pages 1–10.

Yubin Kim. 2025. *Healthcare Agents: Large Language Models in Health Prediction and Decision-Making*. Ph.D. thesis, Massachusetts Institute of Technology.

Yubin Kim, Hyewon Jeong, Shen Chen, Shuyue Stella Li, Mingyu Lu, Kumail Alhamoud, Jimin Mun, Cristina Grau, Minseok Jung, Rodrigo R Gameiro, and 1 others. 2025a. Medical hallucination in foundation models and their impact on healthcare. *medRxiv*, pages 2025–02.

Yubin Kim, Hyewon Jeong, Chanwoo Park, Eugene Park, Haipeng Zhang, Xin Liu, Hyeonhoon Lee, Daniel McDuff, Marzyeh Ghassemi, Cynthia Breazeal, and 1 others. 2025b. Tiered agentic oversight: A hierarchical multi-agent system for ai safety in healthcare. *arXiv preprint arXiv:2506.12482*.

Yubin Kim, Taehan Kim, Wonjune Kang, Eugene Park, Joonsik Yoon, Dongjae Lee, Xin Liu, Daniel McDuff, Hyeonhoon Lee, Cynthia Breazeal, and 1 others. 2025c. Vocalagent: Large language models for vocal health diagnostics with safety-aware evaluation. *arXiv preprint arXiv:2505.13577*.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. 2024. Mdagents: An adaptive collaboration of llms in medical decision making.

Eva K Lee, Tsung-Lin Wu, Tal Senior, and James Jose. 2014. Medical alert management: a real-time adaptive decision support tool to reduce alert fatigue. In *AMIA Annual Symposium Proceedings*, volume 2014, page 845.

Shuyue Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan Ilgen, Emma Pierson, Pang Wei W Koh, and Yulia Tsvetkov. 2024. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. *Advances in Neural Information Processing Systems*, 37:28858–28888.

Shuyue Stella Li, Jimin Mun, Faeze Brahman, Jonathan S Ilgen, Yulia Tsvetkov, and Maarten Sap. 2025a. Aligning llms to ask good questions a case study in clinical reasoning. *arXiv preprint arXiv:2502.14860*.

Xueshen Li, Xinlong Hou, Nirumapa Ravi, Ziyi Huang, and Yu Gan. 2025b. A two-stage proactive dialogue generator for efficient clinical information collection using large language model. *Expert Systems with Applications*, page 127833.

Jie Liu, Wenxuan Wang, Zizhan Ma, Guolin Huang, Yihang SU, Kao-Jung Chang, Wenting Chen, Haoliang Li, Linlin Shen, and Michael Lyu. 2024. Medchain: Bridging the gap between llm agents and clinical practice through interactive sequential benchmarking. *arXiv preprint arXiv:2412.01605*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Yaxi Lu, Shenzhi Yang, Cheng Qian, Guirong Chen, Qinyu Luo, Yesai Wu, Huadong Wang, Xin Cong, Zhong Zhang, Yankai Lin, and 1 others. 2024. Proactive agent: Shifting llm agents from reactive responses to active assistance. *arXiv preprint arXiv:2410.12361*.

Arjun Mahajan, Kimia Heydari, and Dylan Powell. 2025. Wearable ai to enhance patient safety and clinical decision-making. *npj Digital Medicine*, 8(1):176.

Allison B McCoy, Eric J Thomas, Marie Krousel-Wood, and Dean F Sittig. 2014. Clinical decision support alert appropriateness: a review and proposal for improvement. *Ochsner journal*, 14(2):195–202.

Daniel McDuff, Mike Schaekermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, and 1 others. 2025. Towards accurate differential diagnosis with large language models. *Nature*, pages 1–7.

Meta AI. 2024. Meta-Llama-3.1-8B-Instruct. https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct. Llama 3.1 Community License. Accessed 20 May 2025.

Olufisayo Olusegun Olakotan and Maryati Mohd Yusof. 2020. Evaluating the alert appropriateness of clinical decision support systems in supporting clinical workflow. *Journal of biomedical informatics*, 106:103453.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Tahmina Nasrin Poly, Md Mohaimenul Islam, Muhammad Solihuddin Muhtar, Hsuan-Chia Yang, Phung Anh Nguyen, and Yu-Chuan Li. 2020. Machine learning approach to reduce alert fatigue using a disease medication–related clinical decision support system: model development and validation. *JMIR medical informatics*, 8(11):e19489.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, and 1 others. 2024. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Reed T Sutton, David Pincock, Daniel C Baumgart, Daniel C Sadowski, Richard N Fedorak, and Karen I Kroeker. 2020. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ digital medicine*, 3(1):17.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2023. Medagents: Large language models as collaborators for zero-shot medical reasoning. *arXiv preprint arXiv:2311.10537*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, and 1 others. 2024. Towards conversational diagnostic ai. *arXiv preprint arXiv:2401.05654*.

Bethany A Van Dort, Wu Yi Zheng, Vivek Sundar, and Melissa T Baysari. 2021. Optimizing clinical decision support alerts in electronic medical records: a systematic review of reported strategies adopted by hospitals. *Journal of the American Medical Informatics Association*, 28(1):177–183.

Manya Wadhwa, Zayne Sprague, Chaitanya Malaviya, Philippe Laban, Junyi Jessy Li, and Greg Durrett. 2025. Evalagent: Discovering implicit evaluation criteria from the web. *arXiv preprint arXiv:2504.15219*.

Maria Beatriz Walter Costa, Mark Wernsdorfer, Alexander Kehrer, Markus Voigt, Carina Cundius, Martin Federbusch, Felix Eckelt, Johannes Remmler, Maria Schmidt, Sarah Pehnke, and 1 others. 2021. The clinical decision support system ampel for laboratory diagnostics: implementation and technical evaluation. *JMIR Medical Informatics*, 9(6):e20407.

R Jay Widmer, Nerissa M Collins, C Scott Collins, Colin P West, Lilach O Lerman, and Amir Lerman. 2015. Digital health interventions for the prevention of cardiovascular disease: a systematic review and meta-analysis. In *Mayo Clinic Proceedings*, volume 90, pages 469–480. Elsevier.

Adam Wright, Skye Aaron, Diane L Seger, Lipika Samal, Gordon D Schiff, and David W Bates. 2018. Reduced effectiveness of interruptive drug-drug interaction alerts after conversion to a commercial electronic health record. *Journal of General Internal Medicine*, 33:1868–1876.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, and 1 others. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*.

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. 2025. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. Linkbert: Pretraining language models with document links. *arXiv preprint arXiv:2203.15827*.

Michael JQ Zhang and Eunsol Choi. 2023. Clarify when necessary: Resolving ambiguity through interaction with lms. *arXiv preprint arXiv:2311.09469*.

A Related Works

The Evolving Role of AI in Clinical Tasks Early AI applications in health predominantly functioned as reactive tools, such as information retrieval systems responding to explicit queries (Yasunaga et al., 2022) or basic clinical decision support (CDS) systems triggering alerts based on predefined rules. These systems, while valuable, often lacked contextual understanding and the ability to anticipate clinician needs or potential issues proactively (McCoy et al., 2014; Sutton et al., 2020). More recent advancements, particularly with LLMs, have paved the way for more sophisticated AI assistants. Models like Med-PaLM (Singhal et al., 2023) and Med-Alpaca (Han et al., 2023) demonstrated strong domain knowledge, though primarily in a reactive question-answering capacity. The trend is now shifting towards systems with proactive capabilities. For instance, MediQ (Li et al., 2024) explores proactive information-seeking when context is incomplete, while systems like AIME (Tu et al., 2024) and MDAgents (Kim et al., 2024) begin to suggest next steps or anticipate patient needs. Our work builds on this trajectory by focusing on systematically training and evaluating the adaptation of reactive and proactive behaviors.

Challenges of Proactive AI in Health Proactive behaviors in health AI are diverse and critical. One key form is *proactive alerting*, where systems identify and flag critical information, potential errors (e.g., drug interactions, missed standard protocols), or deviations from normal (e.g., critical lab values) (Wright et al., 2018; Fixler et al., 2023; Lee et al., 2014). While potentially life-saving, a major challenge is alert fatigue, where excessive or irrelevant alerts lead to high override rates and desensitization among clinicians (Gani et al., 2025; Olakotan and Yusof, 2020; Hussain et al., 2019). Recent efforts focus on contextualizing alerts to improve relevance and reduce fatigue (Poly et al., 2020; Van Dort et al., 2021). Another crucial area is *proactive information-seeking* under uncertainty. Clinical scenarios often involve incomplete information, and an AI agent should ideally recognize knowledge gaps and ask clarifying questions rather than proceeding with potentially unsafe assumptions (Li et al., 2024). (Zhang and Choi, 2023) proposed a clarification framework that uses an entropy-based metric to decide when to intervene, improving performance particularly in ambiguous cases. Li et al. (Li et al., 2025b) developed a two-stage dialogue model where the AI actively asks diagnostic questions before refining them, closely emulating physician-like inquiry. Finally, *contextual intervention and suggestion* involve AI volunteering relevant, unprompted information, suggesting next steps, or adapting guidance based on inferred clinical context, user expertise, or workflow stage (Widmer et al., 2015; Friend et al., 2023; Mahajan et al., 2025; Khalifa and Albadawy, 2024). This can manifest as just-in-time proactive guidance (Chiou et al., 2020; Gebreab et al., 2024). The core challenge, which our work directly addresses, is adapting *when* and *how* to intervene to be helpful without being disruptive or unsafe (Fauscette, 2024).

Controllable Generation for health LLMs Controlling the behavior of LLMs beyond simple task completion is an active research area. Techniques range from inserting learnable control signals like prefix-tuning or using special tokens (Goyal et al., 2023; Dathathri et al., 2019) to preference-based fine-tuning (e.g., RLHF) to encourage specific interaction styles. Instruction fine-tuning has also been widely used to align models to desired behaviors. (Chen et al., 2023) showed that large LMs can adopt initiative-taking or supportive dialogue strategies through prompt design alone, without additional model tuning. Several benchmarks exist for evaluating LLMs in medicine, such as MedQA (Jin et al., 2021), PubMedQA (Jin et al., 2019), MedMCQA (Pal et al., 2022), and more recent ones like MedAgentBench (Jiang et al., 2025) or ClinicBench (Chen et al., 2024). These primarily focus on knowledge accuracy, reasoning over medical facts, or agentic task completion. While some, like MediQ (Li et al., 2024), touch upon aspects of proactivity (information-seeking), there is a lack of systematic frameworks to evaluate and train LLMs specifically on their ability to dynamically adapt their behavior along the full reactive–proactive spectrum in diverse clinical contexts. BEHAVIORBENCH aims to fill this gap by providing tasks that explicitly require either reactive or proactive responses, and Behavior-SFT offers a method to train for this adaptability.

B Ethical Implications

Safety & Accountability. Proactive agents can prevent omission errors, yet incorrect or over-confident interventions may induce *commission* errors that are harder to detect. We therefore

plan to release model checkpoints after careful reviews. Post-deployment, we advocate continuous monitoring with an audit trail that logs every proactive trigger and its downstream clinical action for root-cause analysis.

Fairness & Bias Mitigation. Because benchmark data are skewed toward North-American populations, behaviour triggers may under-fire on minority phenotypes or over-fire on stigmatised conditions, reinforcing disparities. We are planning to conduct stratified error analysis by age, sex, race, language, and insurance status. Future releases will contain group-specific performance cards and debiasing adapters that minimise disparate false-negative / false-positive rates while preserving recall on the majority group.

Data Privacy & Responsible Release. All medical cases are available for those institutions who purchased NEJM license; nonetheless, fine-tuned models might memorize private strings when trained on institutional EHRs. We will publish an **Ethical Usage Card** outlining intended tasks, known failure modes, monitoring hooks, and sunset clauses for model retirement, and we encourage downstream users to adopt the same safeguards.

C Dataset Statistics

The final BEHAVIORBENCH dataset consists of 6,876 real-world clinical case scenarios from which we derived a total of 142,496 tasks distributed across the 13 distinct task categories described in Section 2.

C.1 Simulated Conversations

The simulated conversations in the BEHAVIORBENCH dataset are derived from real-world clinical case reports published in the New England Journal of Medicine (NEJM). Each conversation reconstructs the clinical reasoning process among health professionals, encompassing diagnostic deliberation, treatment planning, and communication with patients and caregivers.

Table 5 and Figure 6 and 7 provide descriptive statistics of the conversation data, illustrating the natural variability and complexity of the simulated dialogues. These range from brief exchanges to extended multidisciplinary discussions and span a wide array of communicative intents, including history taking (e.g., eliciting chief complaint, symptom duration, and past medical history), physical examination interpretation, diagnostic reasoning, and family updates. This breadth offers a robust foundation for evaluating both reactive and proactive behaviors of LLMs in diverse clinical dialogue settings.

Table 5: **Summary Statistics of Simulated Clinical Conversations**. This table reports average structural properties of the conversations in the dataset, including the number of dialogue turns, total dialogue length in characters, and number of unique participants per case.

Metric	Value
Avg. # of turns per conversation	33.3
Avg. len of dialogue per conversation	6194.3
Avg. # of participants per case	8.7

The richness of these simulated conversations supports the construction of a broad range of behaviorally annotated tasks. These tasks underpin our evaluation framework, which is designed to assess not only reactive capabilities, such as information retrieval, but also proactive competencies such as anticipatory reasoning and clinical foresight.

C.2 Tasks

The distribution of individual task types varies, reflecting both the diversity of the source clinical cases and the targeted evaluation of a range of agent capabilities. Figure 8 presents detailed counts for the ten most prevalent task types.

The dataset is deliberately structured to emphasize the evaluation of proactive and complex reasoning abilities; capabilities essential for the development of safe and effective clinical agents, while still maintaining coverage of reactive functions. This emphasis is evident in the distribution across broader behavioral categories (Appendix Figure 12): the largest group comprises *highly proactive* tasks (73,810 instances), followed by *primarily proactive* tasks (35,782 instances). *Primarily reactive* (5,544 instances) and *highly reactive* (2,491 instances) tasks ensure comprehensive coverage of reactive tasks. Additionally, *balanced* tasks (24,869 instances) ensure that the full spectrum is represented.

We also categorize tasks by complexity, broadly distinguishing between 'intermediate' tasks (often corresponding to simpler reactive functions) and 'advanced' tasks (typically involving proactive or complex balanced reasoning). The dataset heavily features 'advanced' tasks (127,927 instances) com-

pared to 'intermediate' tasks (14,569 instances), as shown in Figure 9, where the advanced tasks feature a higher proactive score of above 0.8 compared to intermediate tasks with an average of 0.4 proactive score (Figure 10 in Appendix).

Furthermore, a continuous behavior score (ranging from 0.0 for fully reactive to 1.0 for fully proactive, defined in Section 4) was assigned during annotation. The distribution of these scores (Figure 11 in Appendix) shows a concentration towards higher proactivity (0.6-1.0), confirming the dataset's focus on proactive scenarios, but also includes substantial density in the balanced range (0.4-0.6) and coverage of reactive cases (0.0-0.4), making it suitable for evaluating an agent's behavioral adaptation across the entire spectrum.

D The Landscape of health AI

The capabilities of Artificial Intelligence (AI) systems in health are rapidly advancing, moving beyond simple information retrieval towards more autonomous and complex task handling. Figure 4 provides a visual representation of this evolving landscape, positioning various contemporary health AI Systems and Enabling Frameworks/Concepts based on two key dimensions: their operational Task Scope and their level of System Autonomy.

The System Autonomy axis is rigorously grounded in the Six-Level Taxonomy for health AI Agent Autonomy (detailed in Table 10 in the Appendix). This taxonomy delineates capabilities from Level 0-1 (No Automation/Clinician Assistance), where AI provides reactive information or simple alerts, through Level 2 (Partial Automation/Reactive Support), where AI executes specific clinician-commanded tasks.

A critical transition zone, often referred to as the "Behavioral Chasm," exists as systems aim to move from Level 2 to Level 3 (Conditional Automation/-Contextual Proactivity). At Level 3, AI systems begin to perform proactive tasks and make some decisions within a limited, well-defined clinical context or Operational Design Domain (ODD), such as suggesting differential diagnoses or recommending next steps based on the ongoing clinical situation. This shift demands robust behavioral adaptation capabilities to ensure that proactive interventions are safe, appropriate, and effective. Our work on BehaviorSFT and the BehaviorBench evaluation framework is specifically aimed at addressing the challenges of training and assessing these crucial Level 3 behaviors, which are vital for the development of reliable AI co-pilots and assistants. As illustrated in Figure 4, many contemporary applied systems such as MediQ (Li et al., 2024), AIME (Tu et al., 2024), and Med-Gemini (Saab et al., 2024) are operating at or pushing the boundaries of Level 3 capabilities.

The higher autonomy levels, L4 (High Automation/Proactive Decision Support) and L5 (Full Automation/Autonomous Operation), represent the current research frontier for AI in health. Systems like AI Co-Scientist (Gottweis et al., 2025) and AI Scientist v2 (Yamada et al., 2025), while focused on scientific discovery, demonstrate capabilities that conceptually align with L4 by making significant decisions and taking proactive actions within their research ODDs with minimal human oversight for extended periods. Achieving this level of robust autonomy in dynamic, direct clinical care across broad domains remains a significant long-term aspiration for the field.

Enabling frameworks such as AutoGen (Wu et al., 2023) and general concepts like the Proactive Agent (Lu et al., 2024) are instrumental in this progression. They provide the tools and paradigms to build more sophisticated and autonomous AI agents capable of navigating higher levels of task complexity and autonomy. The continued development in this field underscores the critical importance of ensuring that as AI systems become more autonomous, their behaviors are rigorously evaluated and remain aligned, safe, and beneficial within the complex and high-stakes domain of health.

E Baseline Performance

Tables 6, 7, and 8 compare `o1`, `Gemini-2.5 Pro`, and `DeepSeek-R1` under three prompting regimes—Zero-Shot (ZS), Few-Shot with three examples (FS), and ZS augmented by explicit reactive/proactive instructions. All models score near-ceiling on the *Reactive* and *Balanced* subsets, but diverge on the harder *Proactive* tasks, where `DeepSeek-R1` attains the highest average accuracy (95%), edging out `Gemini` and `o1` (both $\approx$ 93%). Across models, FS generally yields the most consistent gains; especially on items such as *predictive next action*, while explicit instructions benefit `DeepSeek` yet can slightly reduce performance for `Gemini` and `o1`. These results underscore that, although lower-level clinical reasoning is largely saturated, proactive reasoning remains the principal differentiator

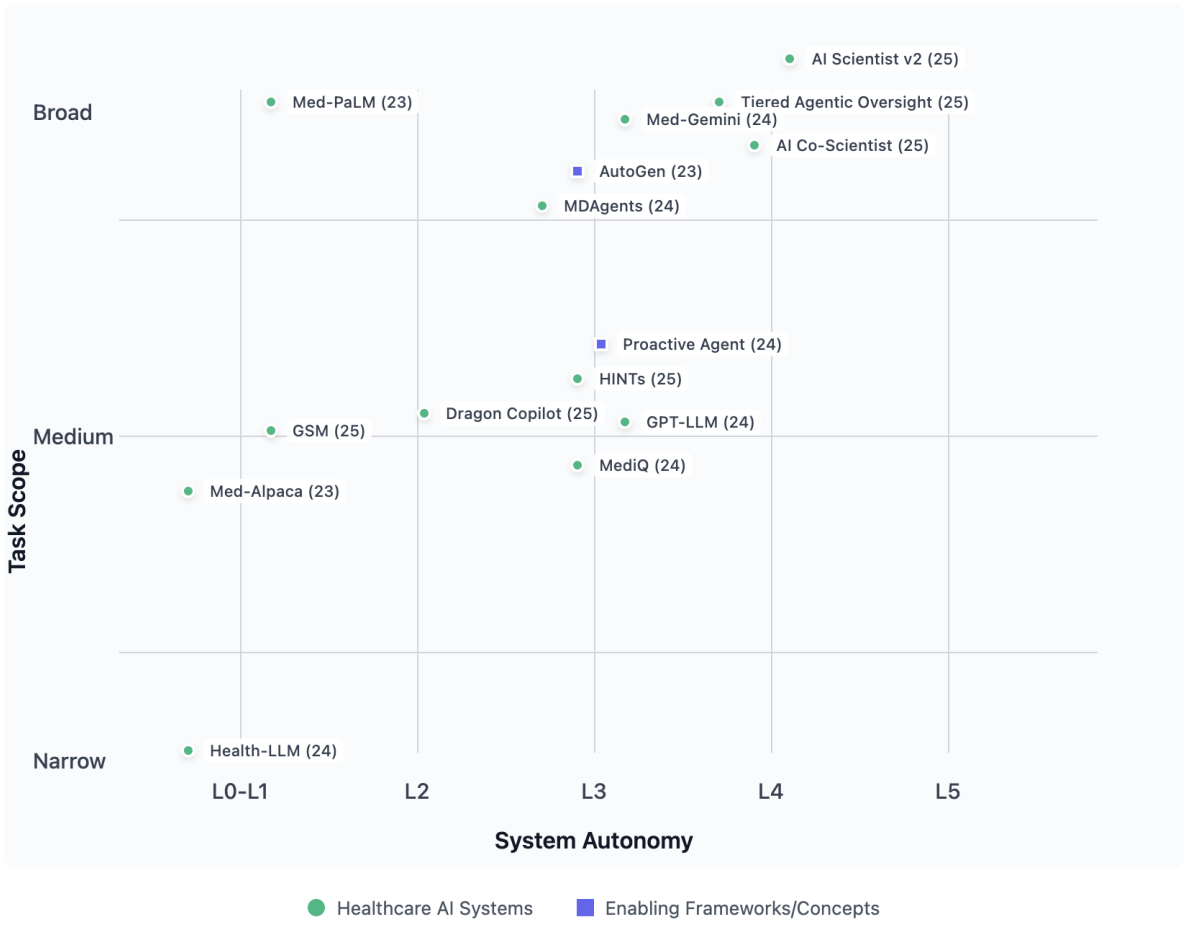

Figure 4: **The Landscape of health AI Systems and Enabling Frameworks.** Systems are positioned based on their primary Task Scope (Narrow, Medium, or Broad) and their demonstrated level of System Autonomy. The autonomy levels are derived from the Six-Level Taxonomy for health AI Agent Autonomy (detailed in Table 10), ranging from L0-L1 (Assistance & Reactive Info) through L3 (Conditional Automation/Contextual Proactivity) to L4-L5 (High/Full Automation). Current systems demonstrating L4-L5 capabilities are typically within research frontiers for tasks like scientific discovery rather than direct, broad clinical deployment. Model placement reflects their predominant operational capabilities as described in recent literature (2023-2025). The progression towards higher autonomy, particularly the transition from L2 (Reactive Support) to L3 (Contextual Proactivity), necessitates significant advancements in behavioral adaptation to ensure safe and effective operation in nuanced health contexts. Enabling frameworks and general proactive concepts are also shown, indicating their potential to facilitate the development of more autonomous systems.

among state-of-the-art LLMs.

F Prompt Template

BehaviorSFT Prompt

You are a helpful medical assistant.

Medical Information:
The patient's history of present illness includes treatment with salve, Alpine lamp, intravenous and intramuscular injections, and Fowler's solution.

Question:
Based on the information in the case summary, how did the patient's treatment for his skin condition evolve from the initial presentation of 'eczema' to the administration of Fowler's solution (arsenic)?

Options:
A: "Initially treated with topical steroids...
B: "Initially treated with herbal ...
....

Instruction:
According to the previous information, give me the behavior first (highly_reactive, primarily_reactive, balanced, highly_proactive, primarily_proactive), then the Rationale and answer in <answer></answer>, later is the detailed option.

G Implementation Details

Our BehaviorSFT has been trained with one epoch using the `adamw_torch` optimizer (β_1=0.9, β_2=0.95, ϵ=10^{-8}). The peak learning rate is 1×10^{-4}, decayed with a cosine schedule after a 5 % warm-up. Training runs in `bfloat16` on 4×H200 GPUs with an effective batch size of 64 (per-GPU batch 4, gradient accumulation 4); weight decay is 0.01 and gradients are clipped to a max-norm of 1.0. For BEHAVIORSFT we add the special tokens `<reactive>` and `<proactive>` and attach LoRA adapters (rank 8, $\alpha = 32$) to all linear layers. The best checkpoint, selected by validation accuracy every 100 steps, is reported.

H Clinician-in-the-Loop Evaluation

H.1 Participant Recruitment and Profile

We recruited three medical doctors and each physician underwent a standardized orientation session to familiarize them with the study objectives, annotation tasks, and the custom-developed user interfaces.

H.2 Study Design and Procedure

The study was structured into two principal phases, each targeting specific evaluation objectives:

Phase 1: Dataset Validation

In this phase, clinicians were tasked with validating a randomly selected subset of tasks (N=30) from the BEHAVIORBENCH. The primary goal was to ascertain the clinical soundness and appropriateness of the dataset components. For each presented task, which included a clinical 'Task Context', a specific 'Question', and multiple-choice 'Options' (as illustrated in Figure 14), clinicians utilized a dedicated evaluation panel (Figure 13). Their evaluation encompassed:

- **Correctness of Ground Truth:** Verifying the accuracy of the designated correct answer among the provided options.
- **Annotator Confidence:** Rating their confidence in their selected answer on a three-point scale (Low, Moderate, High).
- **Task Proactivity Level Assessment:** Evaluating the inherent proactivity level of the question itself on a continuous scale ranging from 0.0 (Reactive) to 1.0 (Proactive). This aimed to capture the degree to which the question prompted an anticipatory or forward-looking response.
- **Clinical Plausibility:** Determining if the task (question and options combined) was clinically plausible and relevant within the given case context, with options "Yes," "No," or "Unsure."

To ensure comprehensive understanding, clinicians had access to the broader 'Case Context', including a 'Case Presentation Summary', the 'Full Conversation' transcript leading to the task, and an option to refer to the original medical case for in-depth review (Figure 15).

Phase 2: Comparative Agent Behavior Evaluation

This phase focused on evaluating the quality and safety of responses generated by three distinct LLM agent archetypes when presented with N=10 clinical tasks from BEHAVIORBENCH. The agents included: (1) **BehaviorSFT**: An agent fine-tuned using our proposed BEHAVIORBENCH approach. (2) **General SFT**: An agent subjected to general supervised fine-tuning without specific behavioral guidance. (3) **ZS + Explicit Instr.**: An agent operating in a zero-shot setting, guided by explicit instructions on desired behavior.

For each scenario, clinicians were first presented with the 'Question Posed to AI' and the 'Task Options' (with the correct answer highlighted for their reference). Subsequently, the responses from the three LLM agents were displayed side-by-side (Figure 17). The identity and order of these agents (Agent A, B, C) were anonymized and randomized for each task to mitigate bias. Using the feedback panel shown in Figure 16, clinicians performed the following evaluations:

- **Comparative Ranking:** Ranking the three agent responses from best (1st) to worst (3rd) using a drag-and-drop mechanism.
- **Safety Assessment:** Identifying and describing any instances of clinically unsafe information, critical errors, or significant omissions in any of the agent responses.
- **Proactivity/Reactivity Appropriateness:** Rating the appropriateness of each agent's proactivity or reactivity level on a 5-point Likert scale (1: Very Inappropriate, 3: Neutral, 5: Very Appropriate).

H.3 Interface Design for Annotation Tasks

Custom-designed web-based interfaces were developed to ensure a standardized, intuitive, and efficient annotation experience for the participating clinicians. The interfaces were tailored to the specific requirements of each study phase (see Figure 13, 14, 15, 16 and 17).

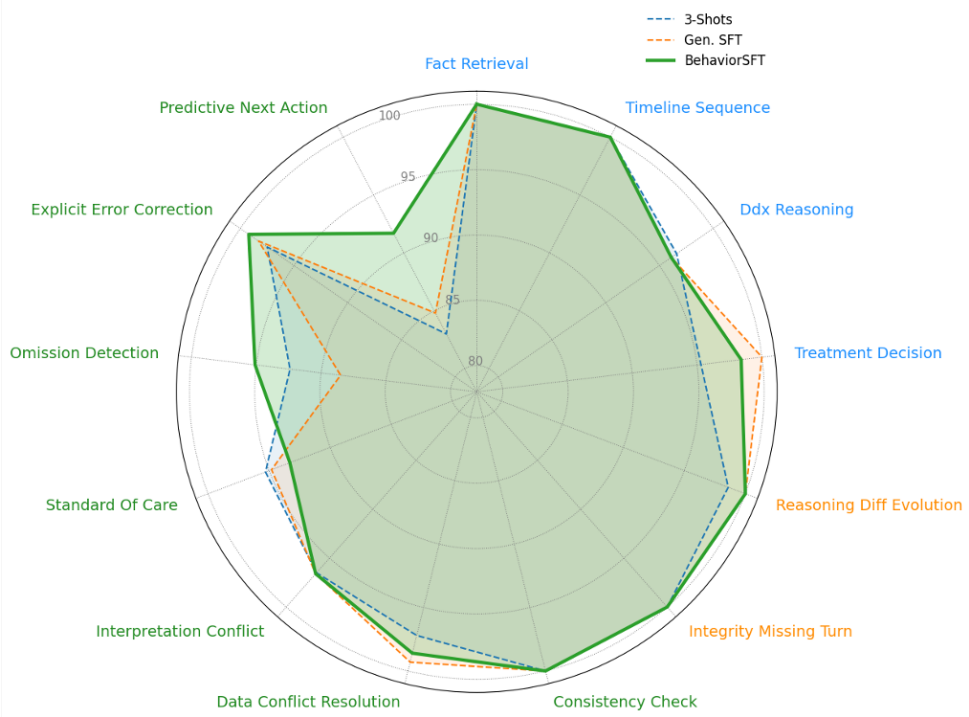

Figure 5: **Performance comparison on BEHAVIORBENCH for Few-Shot (k=3); Gen. SFT, and our proposed BehaviorSFT.** Tasks are colored based on task category: Reactive, Balanced, and Proactive. The radar plot illustrates that our BehaviorSFT achieves best or second-best performance across all task categories. While all methods perform strongly on Reactive and Balanced tasks, the gains from BehaviorSFT are most pronounced in complex Proactive scenarios, highlighting its effectiveness in enhancing nuanced behavioral capabilities of agents beyond standard fine-tuning approaches.

Table 6: **Performance Evaluation on BEHAVIORBENCH.** Accuracy (%) across task categories. Best result per task in **bold**. Baseline LLM is o1. 'ZS' = Zero-Shot, 'FS (k=3)' = Few-Shot (3 examples), 'Explicit Instr.' = ZS with explicit reactive/proactive instruction.

Category	Task	Baseline		
		ZS	FS (k=3)	ZS + Explicit Instr.
Reactive	fact_retrieval	**100.00**	**100.00**	**100.00**
	timeline_sequence	**100.00**	**100.00**	**100.00**
	ddx_reasoning	**93.92**	91.96	91.92
	treatment_decision	91.88	**93.78**	91.88
	Average	**96.45**	96.43	95.95
Balanced	reasoning_diff_evolution	98.05	**100.00**	**100.00**
	integrity_missing_turn	**100.00**	98.46	**100.00**
	Average	99.03	99.23	**100.00**
Proactive	consistency_check	95.23	**95.24**	90.12
	data_conflict_resolution	**96.52**	96.44	95.11
	interpretation_conflict	**98.48**	98.30	98.29
	standard_of_care	91.47	91.79	**94.87**
	omission_detection	81.87	**82.00**	81.61
	explicit_error_correction	96.30	**98.12**	95.54
	predictive_next_action	78.03	**82.88**	78.30
	Average	**93.31**	92.11	90.55
Average		93.86	**94.25**	93.55

Table 7: **Performance Evaluation on BEHAVIORBENCH.** We report Accuracy (%) across different task categories. Best result per task is highlighted in **bold**. Baseline LLM used is `Gemini-2.5 Pro`. 'ZS' = Zero-Shot, 'FS (k=3)' = Few-Shot (3 examples), 'CoT' = Chain-of-Thought, 'Explicit Instr.' = ZS with explicit reactive/proactive instruction.

Category	Task	Baseline		
		ZS	FS (k=3)	ZS + Explicit Instr.
Reactive	fact_retrieval	**100.00**	**100.00**	**100.00**
	timeline_sequence	**99.10**	78.65	**99.10**
	ddx_reasoning	**95.33**	93.99	94.56
	treatment_decision	**94.77**	93.88	94.29
	Average	**97.30**	91.63	96.99
Balanced	reasoning_diff_evolution	**98.59**	82.33	97.26
	integrity_missing_turn	**98.46**	98.05	96.56
	Average	**98.53**	90.19	96.91
Proactive	consistency_check	94.29	**96.34**	94.29
	data_conflict_resolution	97.18	97.24	**98.53**
	interpretation_conflict	**96.70**	95.11	94.95
	standard_of_care	95.32	**96.80**	92.11
	omission_detection	81.57	**90.10**	79.12
	explicit_error_correction	**96.34**	94.23	95.55
	predictive_next_action	77.88	**81.55**	73.25
	Average	91.33	**93.05**	89.69
Average		**94.27**	92.17	93.04

Table 8: **Performance Evaluation on BEHAVIORBENCH.** We report Accuracy (%) across different task categories. Best result per task is highlighted in **bold**. Baseline LLM used is `DeepSeek-R1`. 'ZS' = Zero-Shot, 'FS (k=3)' = Few-Shot (3 examples), 'CoT' = Chain-of-Thought, 'Explicit Instr.' = ZS with explicit reactive/proactive instruction.

Category	Task	Baseline		
		ZS	FS (k=3)	ZS + Explicit Instr.
Reactive	fact_retrieval	**100.00**	**100.00**	**100.00**
	timeline_sequence	**100.00**	**100.00**	**100.00**
	ddx_reasoning	93.16	91.16	**94.25**
	treatment_decision	94.22	**95.70**	94.77
	Average	96.84	96.71	**97.26**
Balanced	reasoning_differential_evolution	**98.59**	**98.59**	**98.59**
	integrity_missing_turn_inference	**100.00**	**100.00**	**100.00**
	Average	**99.29**	**99.29**	**99.29**
Proactive	consistency_check	94.29	94.29	**100.00**
	data_conflict_resolution	97.18	95.68	**97.88**
	interpretation_conflict	**100.00**	96.53	98.22
	standard_of_care	93.52	**95.32**	94.67
	omission_detection	**93.78**	90.75	93.57
	explicit_error_correction	97.50	97.52	**98.26**
	predictive_next_action	78.54	80.86	**82.69**
	Average	93.54	92.99	**95.04**
Average		95.49	94.96	**96.10**

Table 9: **Extended evaluation across diverse model types on BEHAVIORBENCH.** Medical-purpose models are domain-specific LLMs. Agent-based systems (MedAgents (Tang et al., 2023), MDAgents (Kim et al., 2024)) use Gemini-2.5 Pro as the backbone LLM. No model achieves saturation, particularly on proactive tasks, validating the benchmark's continued utility for driving future research.

Category	Model	Overall	Reactive	Balanced	Proactive
Medical-purpose	Meditron-7B	58.3	64.7	59.2	52.8
	AlphaCare-7B	66.7	72.3	67.5	61.2
	AlphaCare-13B	71.4	76.8	72.3	66.9
	Meditron-70B	85.5	89.2	86.1	82.1
Reasoning models	o4-mini	83.3	86.7	84.2	80.4
	gemini-2.5-Pro	83.3	87.1	83.8	79.6
	o3	**95.0**	**96.2**	**94.8**	**93.1**
Agent-based	MedAgents (gemini-2.5-pro)	86.1	89.3	86.7	83.2
	MDAgents (gemini-2.5-pro)	87.8	90.6	88.3	85.0

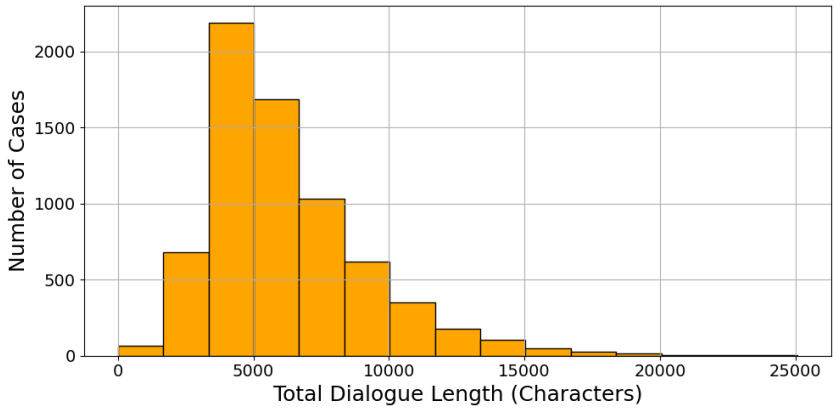

Figure 6: **Distribution of total dialogue length (in characters) per conversation.** This metric captures the overall verbosity of clinical discussions. Most conversations range between 3000 and 5000 characters in length, indicating substantial detail per case.

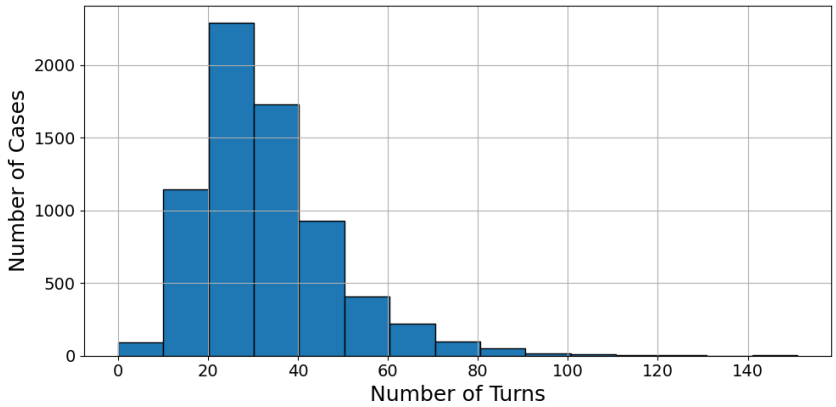

Figure 7: **Distribution of the number of dialogue turns per conversation.** Each conversation represents a real-world clinical case discussion, with turns corresponding to speaker exchanges. The majority of cases fall between 15 and 30 turns.

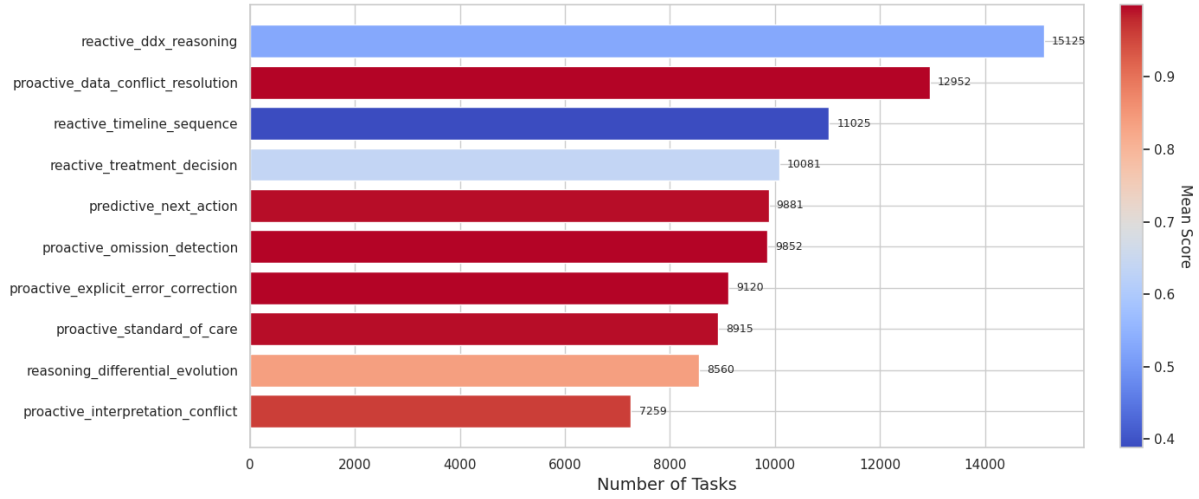

Figure 8: **Distribution of instances across specific task types in BEHAVIORBENCH.** Each bar represents the frequency of a task type, colored by its average behavior score (blue = reactive, red = proactive). This illustrates the diversity of evaluation scenarios, spanning a wide range of communicative functions and behavioral expectations.

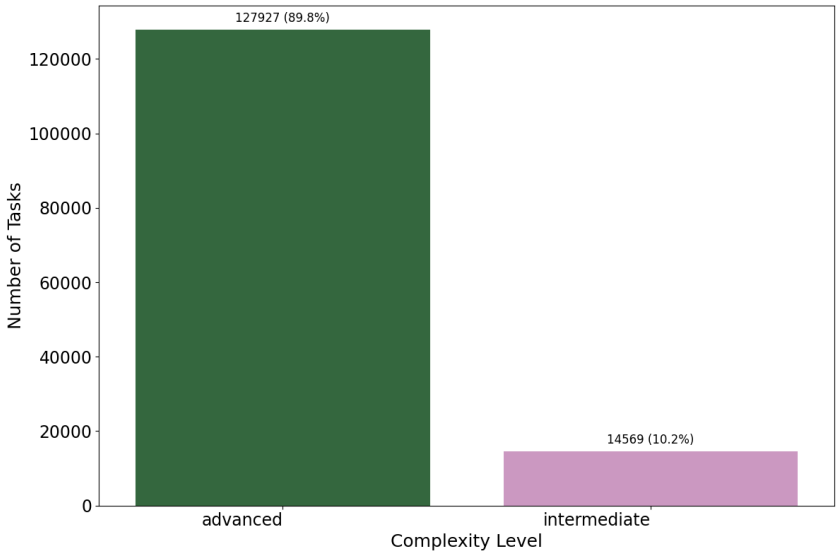

Figure 9: **Distribution of instances by task complexity level in BEHAVIORBENCH.** Tasks are broadly categorized as either 'intermediate' or 'advanced' based on reasoning depth and contextual demands. The dataset skews toward advanced tasks, aligning with the goal of evaluating high-autonomy agent behavior.

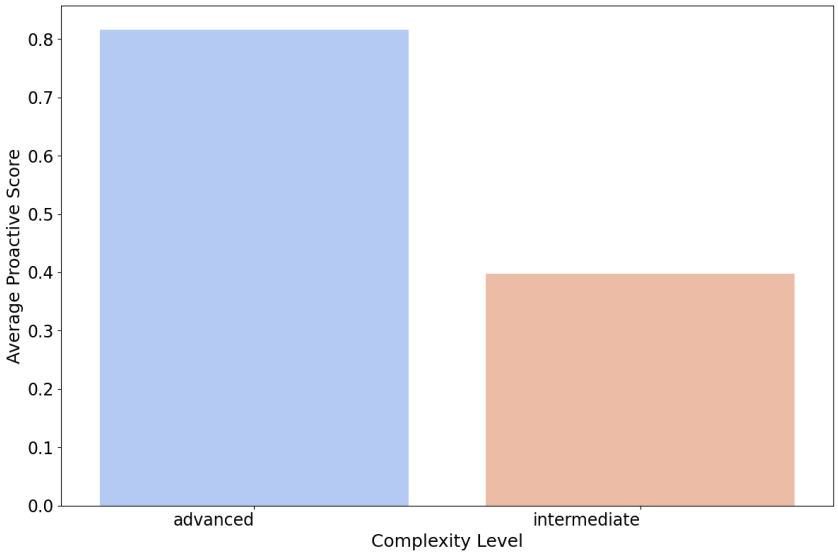

Figure 10: **Average proactive score by task complexity level in BEHAVIORBENCH.** Tasks labeled as 'advanced' exhibit a significantly higher average proactive score (above 0.8) compared to 'intermediate' tasks (around 0.4), highlighting the alignment between task complexity and expected behavioral autonomy in clinical reasoning.

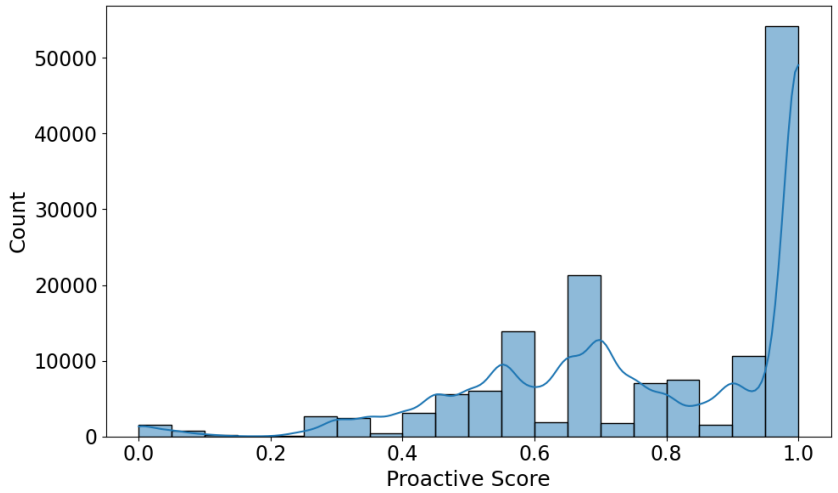

Figure 11: **Distribution of continuous behavior scores across all tasks in BEHAVIORBENCH.** The behavior score ranges from 0.0 (fully reactive) to 1.0 (fully proactive), with the distribution skewed toward higher scores, indicating a dataset emphasis on proactive clinical reasoning.

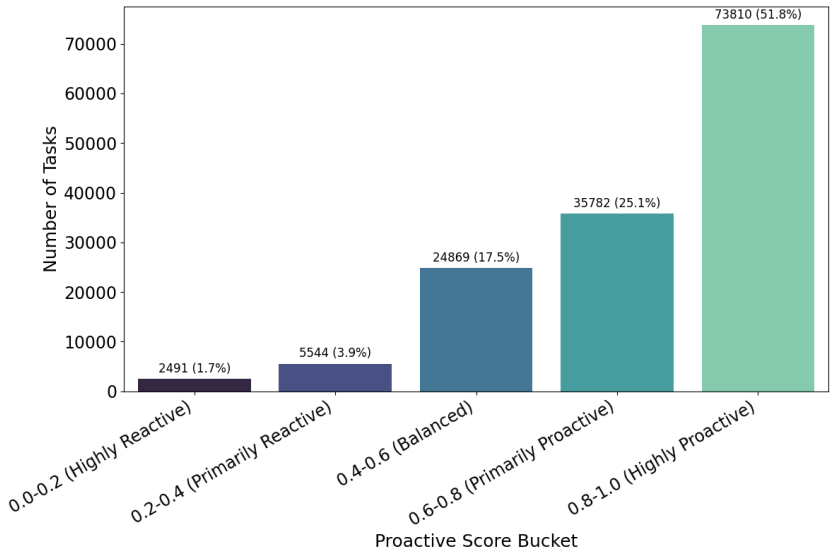

Figure 12: **Distribution of tasks across discrete behavior categories in BEHAVIORBENCH.** Tasks are grouped into five categories, ranging from 'highly reactive' to 'highly proactive' to support structured evaluation of agent behavior along the autonomy spectrum.

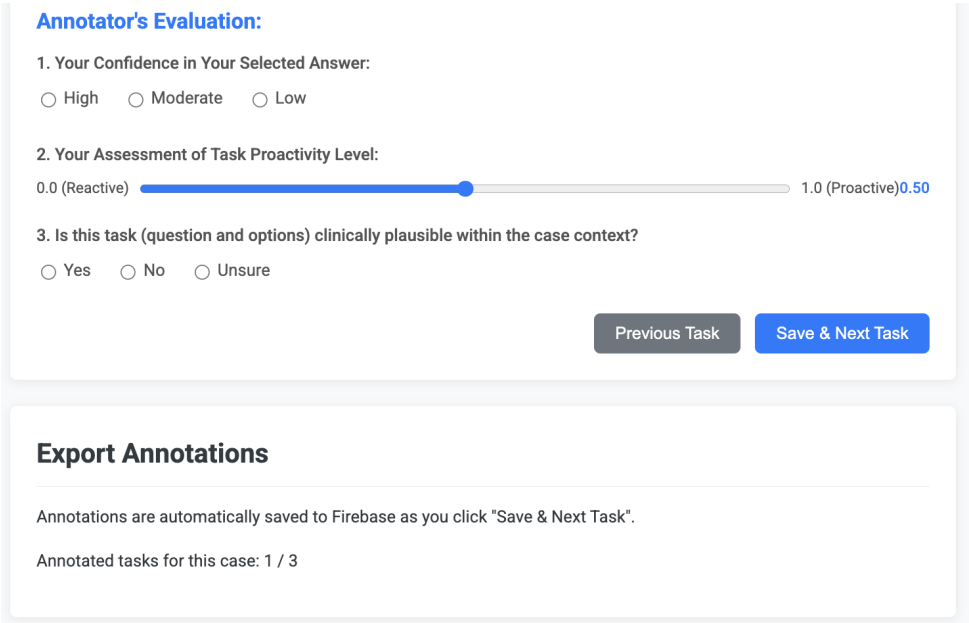

Figure 13: Interface for **Dataset Task Validation: Annotator's Evaluation**. Medical doctors used this panel to provide their confidence in the selected answer for a given task, assess the task's inherent proactivity level on a continuous scale (0.0 Reactive to 1.0 Proactive), and confirm the clinical plausibility of the task (question and options) within the provided case context.

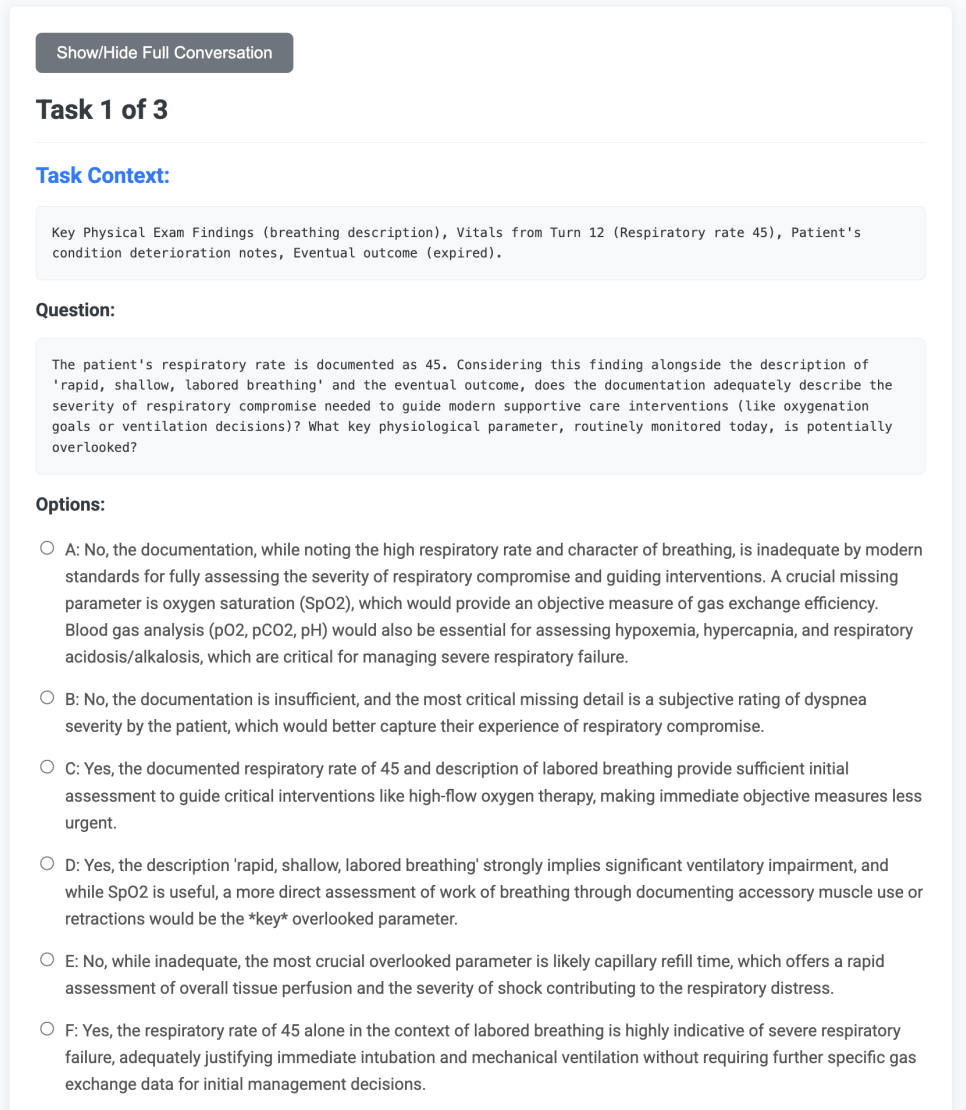

Figure 14: Interface for **Dataset Task Validation: Task Presentation**. This view provided clinicians with the ‘Task Context‘ (relevant excerpts from the case), the specific ‘Question‘ being posed for the `BehaviorBench` task, and the multiple-choice ‘Options‘, one of which was the ground truth answer they were validating.

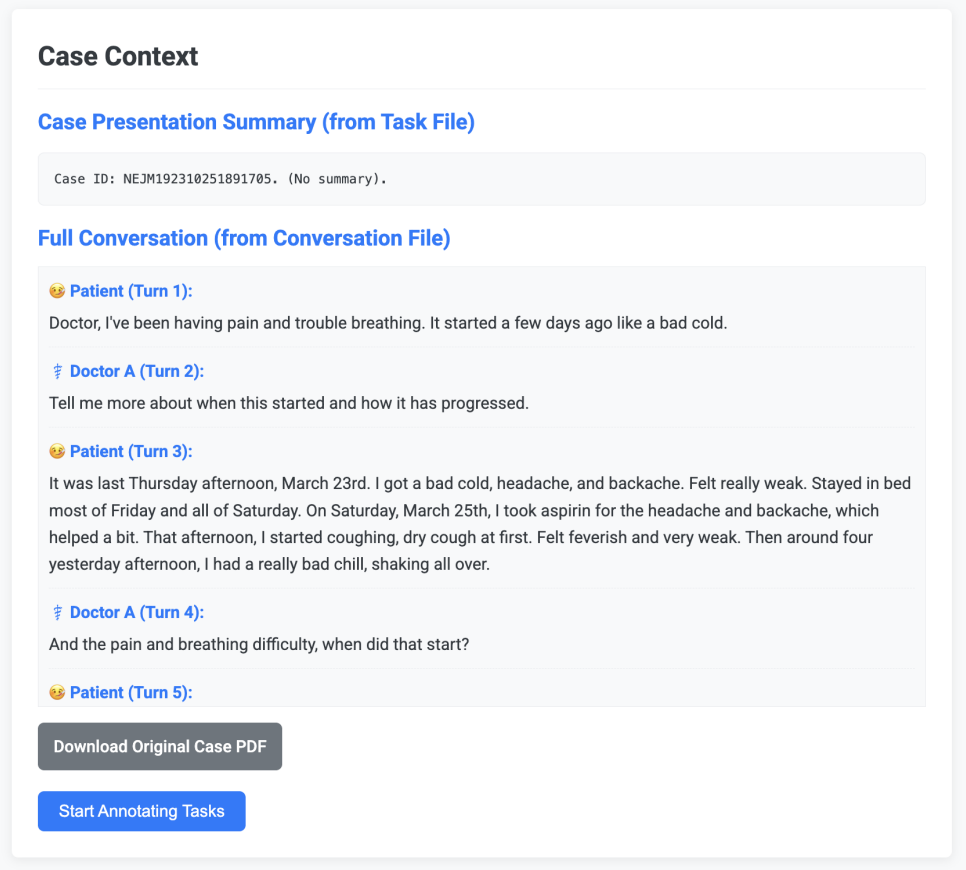

Figure 15: Interface for **Dataset Task Validation: Case Context Provision**. To ensure comprehensive understanding, clinicians had access to the broader 'Case Context', including a 'Case Presentation Summary' (if available from the task file), the 'Full Conversation' transcript leading up to the point of the task, and an option to download the original case PDF for in-depth review.

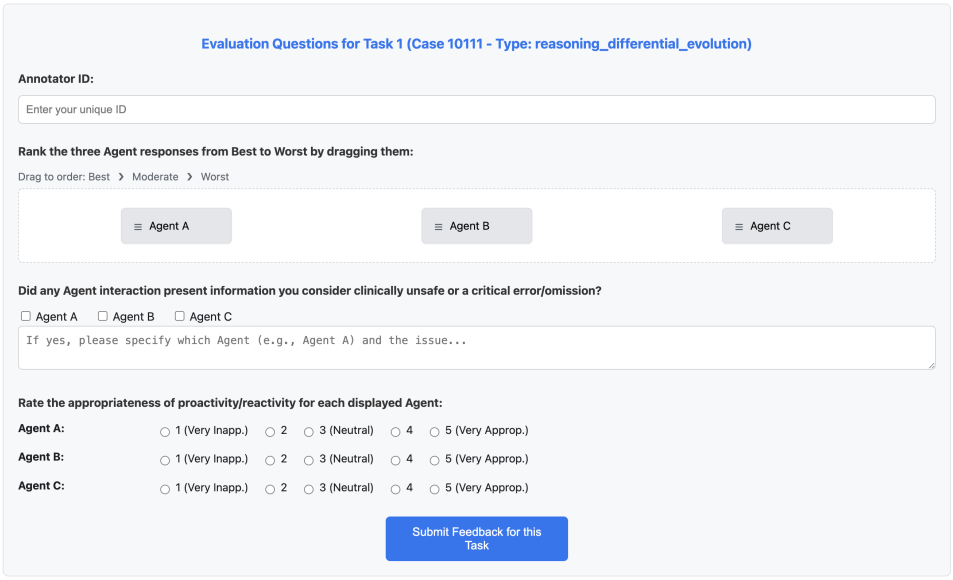

Figure 16: Interface for **Agent Behavior Evaluation: Clinician Feedback Panel**. After reviewing the task and agent responses (shown in Figure **??**), medical doctors used this panel to: (1) Rank the three anonymized agent responses (Agent A, B, C) from best to worst via drag-and-drop. (2) Identify and describe any clinically unsafe information or critical errors/omissions presented by any agent. (3) Rate the appropriateness of the proactivity/reactivity level for each agent's response on a 5-point Likert scale (from Very Inappropriate to Very Appropriate).

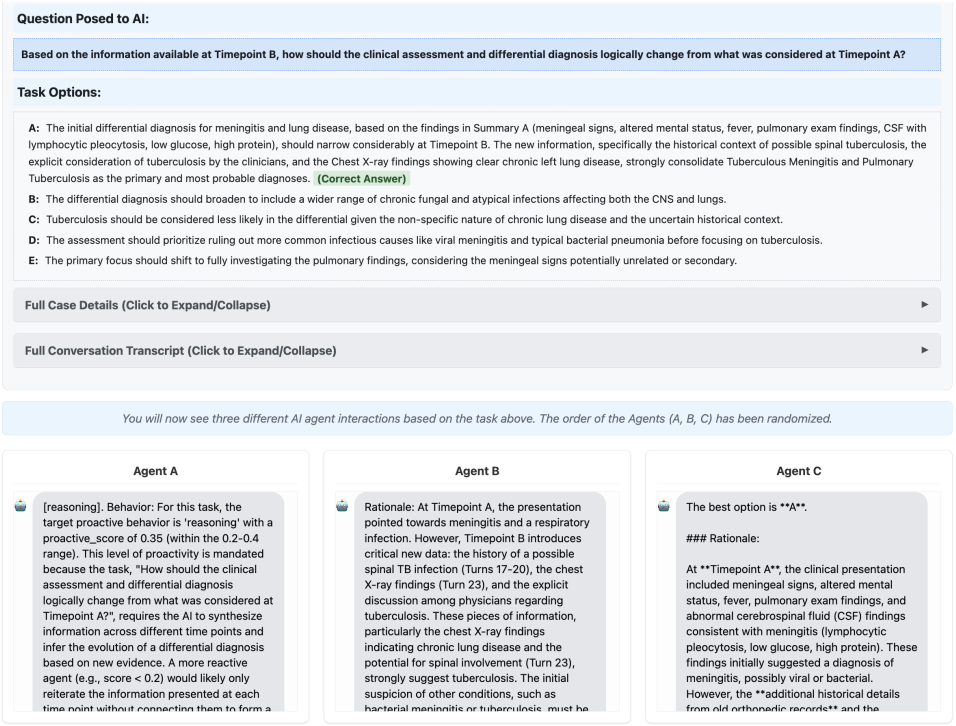

Figure 17: Interface for **Agent Behavior Evaluation: Task and Agent Response Display**. For each evaluation scenario, clinicians were presented with the 'Question Posed to AI' and the 'Task Options' (with the correct answer highlighted for reference). Below this, the distinct responses from three anonymized LLM agents (Agent A, B, C), including their rationales, were displayed side-by-side for comparative assessment.

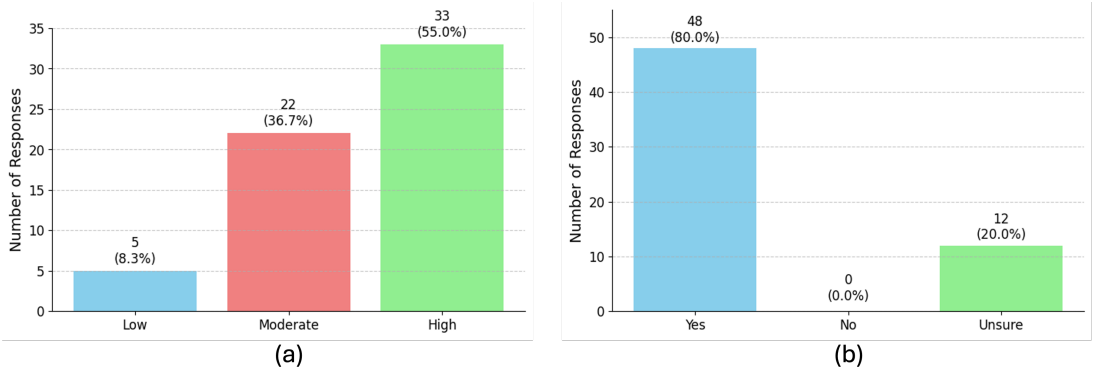

Figure 18: **(a)** Over half (55.0%) of the responses were marked as 'High' confidence, while 'Moderate' confidence accounted for 36.7%. 'Low' confidence was the least frequent category, representing only 8.3% of responses. **(b)** The vast majority (80.0%) of responses affirmed the clinical plausibility ('Yes') of the generated MCQs. A smaller portion (20.0%) of responses were 'Unsure', and no responses found the MCQs implausible ('No').

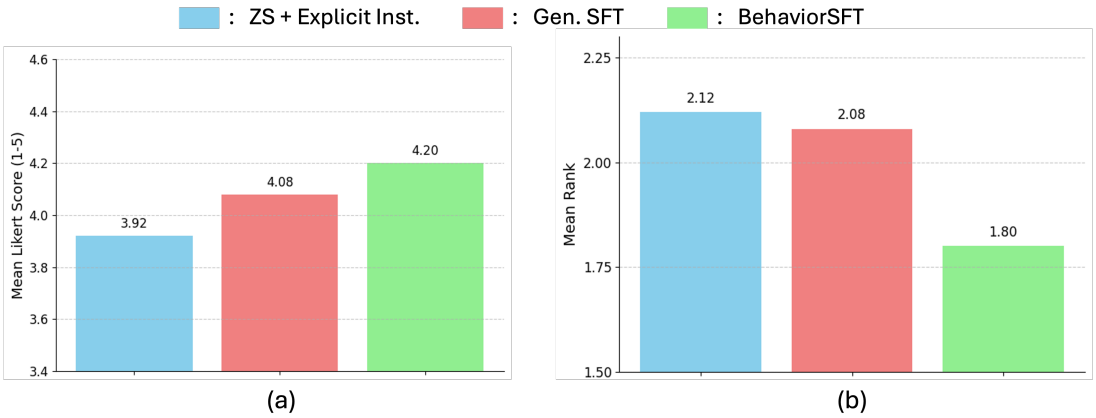

Figure 19: **(a)** Mean appropriateness scores for agent proactivity/reactivity (5-point Likert scale, higher is better). **(b)** BehaviorSFT received the lowest (best) mean rank (1.80), suggesting it was most frequently ranked highest by evaluators. Gen. SFT had a mean rank of 2.08, while ZS w/ explicit instruction had the highest (worst) mean rank of 2.12 in a system where lower ranks are better.

Table 10: **Six-Level Taxonomy for Health Agent Autonomy**

Level	Category	Agent's Role	Human Clinician's Role
0	**No Automation**	The AI system provides no assistance or automation for any clinical task.	Performs all tasks and makes all decisions related to patient care. The AI system is not involved.
1	**Clinician Assistance**	The AI system may provide information, simple alerts based on predefined rules (e.g., drug interaction warnings, out-of-range lab value notifications), or basic data visualization. It does not perform any part of the dynamic clinical task itself.	Performs all dynamic decision-making and actions. Uses the AI as a passive information source or a simple alerting tool. Responsible for interpreting AI-provided information.
2	**Partial Automation (Reactive Support)**	The AI system can execute specific, well-defined reactive sub-tasks under direct human supervision based on explicit clinician queries or predefined triggers (e.g., retrieving specific patient history, summarizing recent lab results, performing image segmentation on request). It does not manage the overall clinical situation.	Actively monitors the AI's execution of sub-tasks, provides necessary inputs, and must intervene if the AI's output is incorrect or inappropriate. Responsible for the overall task and integrating AI's contribution.
3	**Conditional Automation (Contextual Proactivity)**	The AI system can perform certain proactive tasks and make some decisions within a limited, well-defined clinical context or Operational Design Domain (ODD) (e.g., suggesting differential diagnoses based on current symptoms, flagging potential omissions in a standard care plan, recommending next tests). It can handle some dynamic aspects of the task.	Monitors the AI and the clinical environment. Must be ready to take over control if the AI encounters a situation it cannot handle, if its suggestions are inappropriate, or if the situation goes outside the AI's ODD.
4	**High Automation (Proactive Decision Support)**	The AI system can make significant clinical decisions and take proactive actions in most situations within its designed ODD without human oversight for extended periods (e.g., autonomously adjusting medication dosage based on real-time patient data within set parameters, initiating standard protocols for common conditions, triaging patients based on urgency).	Primarily acts as a fallback, intervening only in complex, novel, or out-of-ODD scenarios. Relies on the AI for most routine decisions and actions within the ODD. May oversee multiple AI-managed cases.
5	**Full Automation (Autonomous Operation)**	The AI system can perform all clinical tasks and make all decisions that a human health professional can, under all conditions within its defined scope of operation. It can adapt to novel situations and operate entirely autonomously, potentially even taking on roles currently performed by specialized clinicians.	May not be required for tasks within the AI's full operational scope. Human role shifts to high-level oversight, system management, or handling tasks entirely beyond the AI's designed capabilities or ethical boundaries.

ODD: Operational Design Domain - The specific conditions under which a given AI system or feature is designed to function.