

# Beyond the Individual: Virtualizing Multi-Disciplinary Reasoning for Clinical Intake via Collaborative Agents

Anonymous ACL submission

## Abstract

The initial outpatient consultation is critical for clinical decision-making, yet it is often conducted by a single physician under time pressure, making it prone to cognitive biases and incomplete evidence capture. Although the Multi-Disciplinary Team (MDT) reduces these risks, they are costly and difficult to scale to real-time intake. We propose Aegle, a synchronous virtual MDT framework that brings MDT-level reasoning to outpatient consultations via a graph-based multi-agent architecture. Aegle formalizes the consultation state using a structured SOAP representation, separating evidence collection from diagnostic reasoning to improve traceability and bias control. An orchestrator dynamically activates specialist agents, which perform decoupled parallel reasoning and are subsequently integrated by an aggregator into a coherent clinical note. Experiments on ClinicalBench and a real-world RAPID-IPN dataset across 24 departments and 53 metrics show that Aegle consistently outperforms state-of-the-art proprietary and open-source models in documentation quality and consultation capability.

## 1 Introduction

The trajectory of clinical care is fundamentally established during the initial consultation (Starfield et al., 2005). In this pivotal phase, a physician must transmute a patient’s unstructured narrative of symptoms and concerns into a structured medical record, crystallizing it as the Initial Progress Note (IPN) in the SOAP (Subjective, Objective, Assessment, Plan) format. This document serves as more than a mere administrative summary; it is the cornerstone for downstream diagnostic decisions and treatment planning (Krishna et al., 2021). Consequently, the comprehensiveness and accuracy of this intake process are paramount to effective healthcare delivery, serving as the bedrock upon which the entire clinical pathway rests.

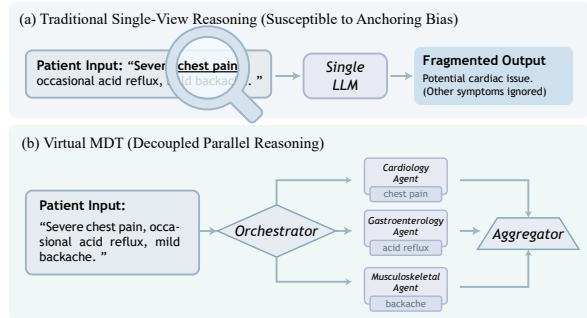


Figure 1: Single-view vs. virtual MDT reasoning for clinical intake. (a) A single LLM is prone to anchoring, over-focusing on salient symptoms, and producing fragmented notes. (b) A virtual MDT enables decoupled parallel specialist reasoning coordinated by an Orchestrator and integrated by an Aggregator, improving evidence coverage and coherence.

However, achieving high-quality intake in routine practice is fraught with cognitive and systemic challenges (You et al., 2025). The traditional process typically relies on a single physician operating under significant time constraints. When formulating diagnoses while simultaneously engaging in empathetic dialogue, individual providers are susceptible to anchoring bias (Croskerry, 2013), fixating on prominent symptoms while overlooking subtler, yet critical, diagnostic clues. As illustrated in Fig. 1(a), this “single-view” setting narrows the exploration of the diagnostic space and can lead to fragmented evidence capture in the resulting note. This is not merely a matter of physician competence but a fundamental limit of human cognitive bandwidth when processing high-entropy patient narratives under time pressure.

To address complex cases where a single perspective is insufficient, medical practice traditionally turns to a Multi-Disciplinary Team (MDT) (Taylor et al., 2010). By aggregating specialists from diverse fields, MDT enables parallel and complementary reasoning across different clin-

ical perspectives, mitigating the anchoring effects inherent to single-view decision making (Fig. 1(b)). While this collaborative model significantly reduces the risk of oversight, it is inherently resource-intensive, asynchronous, and difficult to scale. Organizing a team of human experts for every routine outpatient consultation is logistically impractical. Thus, a critical gap remains: *how can we transpose the systematic depth of MDT-level reasoning to the widespread, real-time outpatient intake phase without prohibitive resource costs?*

In this paper, we bridge this gap by proposing **Aegle**, a multi-agent framework that virtualizes the MDT paradigm. Rather than relying on a single Large Language Model (LLM) or a static chain of agents, Aegle introduces a novel computational architecture for medical inquiry. We posit that the essence of effective collaboration lies in *decoupled parallel reasoning*, where distinct specialist agents analyze the case from their unique domain perspectives without interference, followed by a semantic aggregation phase. Furthermore, to address the efficiency issues common in multi-agent systems, we implement a dynamic topology controlled by a meta-cognitive orchestrator. This allows the system to adaptively scale its reasoning network based on the real-time completeness of the clinical documentation.

Our contributions are summarized as follows:

- We propose **Aegle**, a Synchronous Virtual MDT framework that leverages decoupled parallel reasoning to transcend physical resource constraints. This paradigm transposes systematic, inpatient-level diagnostic depth into real-time outpatient inquiries, significantly enhancing robustness while mitigating single-view cognitive biases.
- We propose a **State-Aware Dynamic Topology** that aligns multi-agent collaboration with the evolving clinical document. By implementing on-demand specialist activation, this mechanism dynamically constructs reasoning paths tailored to case-specific ambiguity, thereby maximizing the diagnostic signal-to-noise ratio and ensuring high-density information gathering.
- We conduct a comprehensive evaluation across **24 clinical departments** using **53 fine-grained metrics**. Empirical results demonstrate Aegle’s superiority over state-of-the-art

baselines in diagnostic accuracy and documentation quality, establishing a robust benchmark for next-generation clinical AI assistants.

## 2 Related Works 119

### 2.1 LLMs for Clinical Consultation and Documentation 120

LLMs have shown promise in clinical workflow optimization tasks such as clinical documentation support and conversational assistance during patient intake (Zhou et al., 2025a). In the realm of documentation, models function as semantic compressors, transforming unstructured dialogues into standardized formats like SOAP notes (Krishna et al., 2021). While models such as Med-PaLM 2 have achieved accuracy comparable to human scribes in summarizing static records (Singhal et al., 2025), they exhibit significant fragility in temporal reasoning. Specifically, when summarizing longitudinal patient trajectories, these models often succumb to the “lost-in-the-middle” effect, failing to accurately distinguish between historical ailments and current presenting symptoms, thereby compromising the integrity of the medical record (Kruse et al., 2025; Zeng et al., 2025).

Conversely, in interactive consultation, frameworks such as AMIE (Tu et al., 2025) and Healthcare Agent (Ren et al., 2025) have attempted to simulate the diagnostic inquiry process. Despite their conversational fluency, a critical limitation persists: these monolithic systems largely operate as passive information receivers (Zhou et al., 2025b). Rather than executing proactively asking rule-out questions to narrow the differential diagnosis space, they tend to hallucinate details or prematurely commit to a diagnosis based on incomplete user input (Qiu et al., 2025). This passivity reveals a fundamental misalignment with real-world intake, where the core challenge lies not merely in processing available text, but in the strategic elicitation of missing evidence (Brooks et al., 2024).

### 2.2 Multi-Agent Systems in Medical Reasoning 156

To transcend the cognitive bounds of single-model architectures, Multi-Agent Systems (MAS) have emerged as a computational analogue to the clinical MDT (Han et al., 2025). By assigning specialized roles such as oncologists, radiologists, and pathologists to distinct agents, frameworks like MedAgents (Tang et al., 2024) and MAC (Chen et al.,

2025) leverage dialectical debate to decompose complex diagnostic tasks. More advanced architectures, such as MDAgents, introduce an adaptive topology that dynamically structures collaboration based on the perceived medical complexity of the case, thereby optimizing the trade-off between accuracy and computational cost (Kim et al., 2024). Furthermore, frameworks such as ClinicalLab have demonstrated the utility of agentic collaboration in managing multi-departmental diagnostics, simulating the referral and consultation dynamics of a physical hospital (Yan et al., 2025).

However, the “black-box” interaction between agents introduces varying degrees of collaborative failure modes. A recent large-scale audit of medical MAS reveals that agentic collaboration can lead to “flawed consensus” where agents reinforce each other’s biases, and the suppression of correct minority opinions during the voting process (Gu et al., 2025). Additionally, the interplay of multiple probabilistic models creates a problem of “compound opacity”, making it exponentially difficult to trace the provenance of a clinical error (Salehi et al., 2025). These vulnerabilities highlight that while MAS can broaden the hypothesis space, they require rigid structural constraints to prevent unanchored speculation, a gap that our proposed framework specifically addresses.

### 3 Methodology

We propose **Aegle**, a multi-agent consultation framework designed to virtualize the cognitive benefits of MDT collaboration during early-stage patient encounters. As illustrated in Fig. 2, Aegle integrates structured, virtualized MDT-style collaboration directly into the consultation workflow. By coordinating multiple specialized agents during information gathering, the framework aims to surface overlooked considerations earlier and to reduce bias arising from single-perspective reasoning.

Built upon DeepSeek-V3.2 (DeepSeek-AI et al., 2025), Aegle instantiates a constrained graph-based agentic topology. Agent interactions are governed by explicit state representations and execution protocols, enabling controllable, transparent, and bias-aware clinical dialogue.

#### 3.1 Structured Clinical State

To ground multi-agent collaboration in established clinical practice, Aegle formalizes the consultation

state using the canonical SOAP schema. We denote the clinical state at turn  $t$  as  $\mathcal{S}_t$ . Beyond its role as a documentation standard, SOAP provides a cognitive structure that explicitly separates evidence collection from diagnostic interpretation, thereby supporting bias-aware reasoning.

We decompose  $\mathcal{S}_t$  into two functionally distinct components:

- **Case Features ( $\mathcal{F}$ )**. Corresponding to the *Subjective* and *Objective* sections of SOAP,  $\mathcal{F}$  serves as an incremental repository of factual evidence. It accumulates verifiable patient information throughout the consultation, including Basic Information, History of Present Illness, Past Medical History, Physical Examination, and Auxiliary Examination results.
- **Diagnosis and Plan ( $\mathcal{P}$ )**. Corresponding to the *Assessment* and *Plan* sections,  $\mathcal{P}$  represents the analytical output of the consultation. It includes the preliminary diagnosis, diagnostic reasoning, and treatment plan, all of which are derived exclusively from the finalized case features in  $\mathcal{F}$ .

The structured state is defined as  $\mathcal{S}_t = [\mathcal{F}_t, \mathcal{P}_t]$  and functions as a shared blackboard accessible to all agents. Aegle enforces a unidirectional dependency from  $\mathcal{F}$  to  $\mathcal{P}$  such that diagnostic and planning components may only be generated after evidence stabilization. This constraint explicitly links clinical conclusions to accumulated evidence, ensuring traceability and mitigating premature commitment to unsupported hypotheses.

#### 3.2 Multi-Agent Graph Topology

Aegle operationalizes virtual MDT collaboration through a dynamic multi-agent graph topology composed of three types of nodes, each fulfilling a distinct role in the consultation workflow.

**Orchestrator.** The Orchestrator acts as a routing and coordination policy  $\pi_{\text{orch}}$  that governs agent activation. It does not perform medical reasoning itself. Instead, it allocates computational attention by selecting a subset of specialist agents based on the evolving consultation context:

$$A_{\text{sub}, \iota} = \pi_{\text{orch}}(\mathcal{H}_t, \mathcal{F}_t), \quad A_{\text{sub}} \subseteq \mathcal{A}_{\text{total}}, \quad (1)$$

where  $\mathcal{H}_t$  denotes the dialogue history and  $\iota$  specifies context-dependent task instructions. This selective activation mechanism mirrors real-world

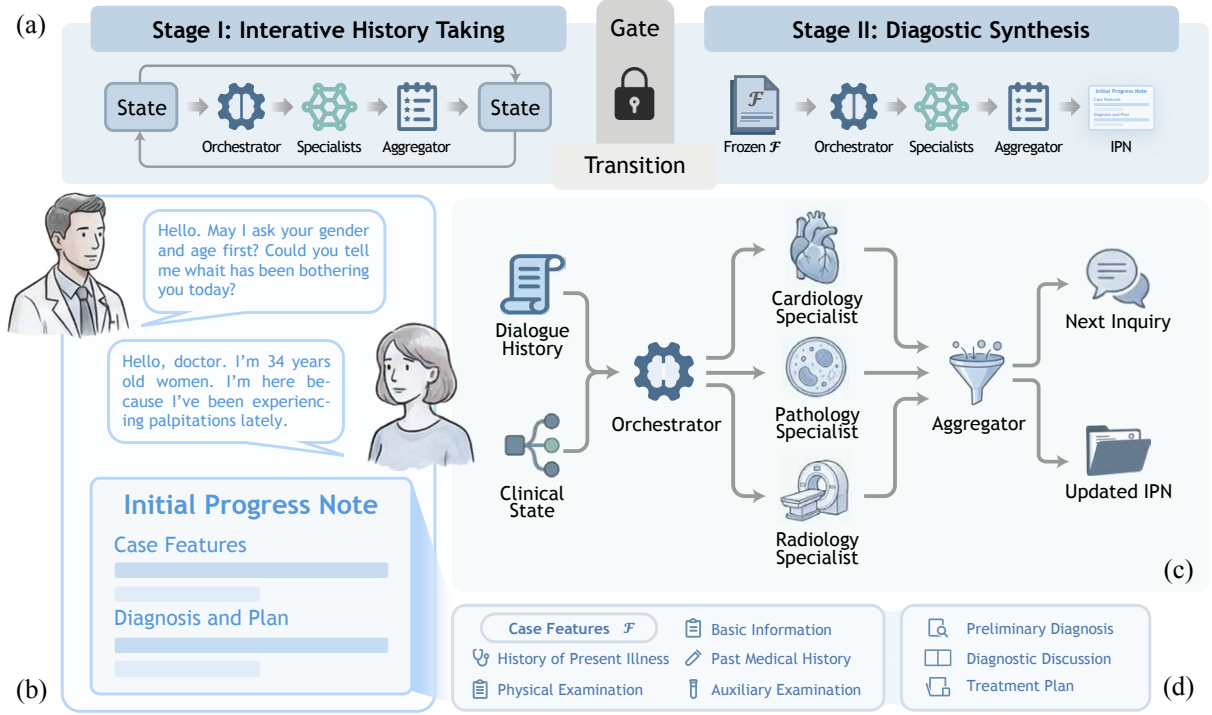


Figure 2: Overview of the Aegle framework. (a) A two-stage consultation workflow consisting of iterative history taking followed by diagnostic synthesis after freezing the case feature set  $\mathcal{F}$ . (b) An evolving draft Integrated Patient Note (IPN) that is incrementally updated throughout the consultation. (c) Dynamic multi-agent collaboration, where a context-aware Orchestrator activates relevant specialist agents and an Aggregator integrates their outputs to update the clinical state and generate the next inquiry. (d) Structured clinical state  $\mathcal{S}_t = [\mathcal{F}_t, \mathcal{P}_t]$  separating evidentiary features from diagnostic and planning components.

MDT practice by engaging specialized expertise only when warranted by the available evidence, thereby avoiding unnecessary or premature expert involvement during early-stage information gathering.

**Specialist Agents.** Each specialist agent operates as an independent domain expert, analyzing the clinical state from a distinct medical perspective. Specialists are executed in parallel and generate proposed updates to the clinical state in isolation. This decoupled architecture preserves hypothesis diversity by construction and delays consensus formation, reflecting the cognitive advantage of independent expert opinions in MDT discussions.

**Aggregator.** The Aggregator  $\pi_{\text{agg}}$  serves as the interface between internal agent reasoning and patient-facing communication. It follows a write-then-speak protocol. First, it validates and integrates specialist proposals to update the structured clinical state:

$$\mathcal{S}_{t+1} = \pi_{\text{agg}}^{\text{write}} \left( \mathcal{S}_t, \{ \Delta \mathcal{S}_t^{(a)} \}_{a \in A_{\text{sub}}} \right). \quad (2)$$

Subsequently, it generates the patient-facing utterance conditioned solely on the updated state:

$$u_{t+1} = \pi_{\text{agg}}^{\text{speak}}(\mathcal{S}_{t+1}). \quad (3)$$

This separation ensures internal consistency and technical precision of the medical record while maintaining clear and empathetic communication with the patient.

### 3.3 Sequential Clinical Execution

Building upon the structured clinical state and defined agent roles, Aegle executes consultations through a two-stage finite state machine. This temporal structure enforces a strict separation between evidence acquisition and diagnostic reasoning, serving as an explicit bias-control mechanism.

**Stage I: Iterative History Taking.** As shown in Fig. 3, the consultation begins with iterative history taking. The Orchestrator activates relevant specialist agents based on patient responses and the current completeness of  $\mathcal{F}_t$ . Each specialist examines the updated clinical state from its domain perspective and proposes follow-up questions to-

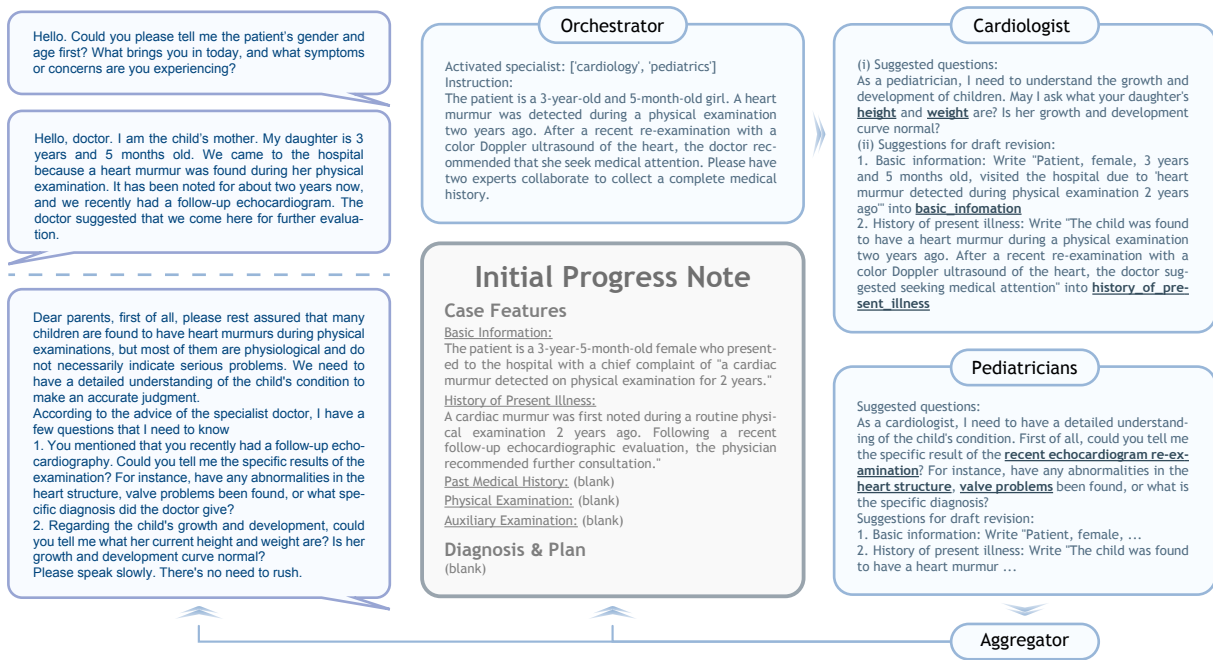


Figure 3: Stage I of Aegle, iterative history taking, illustrated with a pediatric heart murmur case. Patient responses are incorporated into the structured SOAP state and analyzed by multiple specialist agents in parallel. Each specialist proposes follow-up questions and targeted updates to case features. The Aggregator integrates these suggestions and generates the next patient-facing inquiry under a write-then-speak protocol.

gether with evidence-centric revisions to the draft Integrated Patient Note.

The Aggregator integrates these parallel proposals, updates the case features in  $\mathcal{F}$ , and generates the next consultation question. This process continues until all mandatory fields in  $\mathcal{F}$  are either populated or explicitly marked as unavailable by the patient, ensuring that downstream diagnostic reasoning is grounded in sufficient evidence.

**Stage II: Diagnostic Synthesis.** Once  $\mathcal{F}$  is finalized, it is frozen to prevent further modification, and the system transitions deterministically to diagnostic synthesis. In this stage, the Orchestrator commissions specialist agents to perform independent diagnostic reasoning based on the same fixed evidentiary substrate. Specialists propose diagnostic hypotheses and treatment considerations without introducing new inquiries.

The Aggregator then integrates these heterogeneous perspectives to produce the final diagnosis and plan  $\mathcal{P}$ , resolving inconsistencies and generating a complete and coherent SOAP note. This staged execution ensures that diagnostic conclusions are derived exclusively from stabilized evidence, reinforcing traceability and reducing bias induced by early hypothesis fixation.

## 4 Experiments

### 4.1 Datasets

To evaluate Aegle’s performance across diverse clinical scenarios, we utilize two distinct datasets:

**ClinicalBench.** We employ ClinicalBench (Yan et al., 2025), a comprehensive end-to-end benchmark derived from de-identified electronic health records (EHRs) of top-tier Grade 3A hospitals in China. It contains 1,500 cases covering 24 clinical departments and 150 diseases. Crucially, it enforces a strict data-leakage-free protocol and supports open-ended generation tasks, simulating the whole trajectory from triage to treatment planning.

**RAPID-IPN.** To evaluate complex differential diagnosis, we curated the Real-world Abdominal Pain Integrated Diagnostic-pathway Initial Progress Note (RAPID-IPN) dataset from a top-tier Grade 3A hospital in Eastern China. Spanning from 2018 to 2024, this cohort comprises 322 patients with abdominal pain across 12 departments, encompassing internal medicine (e.g., Cardiology, Gastroenterology) and surgery (e.g., Hepatobiliary Surgery). A rigorous review protocol by three senior physicians (>5 years of experience) ensured that the standardized SOAP notes and treatment plans strictly aligned with patients’

Model	ClinicalBench				RAPID-IPN			
	IDEA	SOAP	READ	chrF++	IDEA	SOAP	READ	chrF++
HuatuogPT-o1-7B <sup>†‡</sup>	46.54 ± 11.16	37.98 ± 14.11	68.66 ± 6.30	9.62 ± 1.98	41.89 ± 7.69	40.28 ± 12.96	68.58 ± 3.00	6.66 ± 1.23
Qwen3-8B-think <sup>†</sup>	43.18 ± 9.22	27.66 ± 10.69	66.74 ± 5.48	10.19 ± 2.11	39.93 ± 8.99	28.92 ± 13.42	66.14 ± 5.67	6.89 ± 1.73
Baichuan-M2-32B <sup>†‡</sup>	38.02 ± 9.05	25.58 ± 10.63	64.00 ± 6.93	11.46 ± 2.45	39.40 ± 9.89	31.54 ± 12.25	65.42 ± 5.81	9.24 ± 2.73
Lingshu-32B <sup>†‡</sup>	50.50 ± 10.62	39.76 ± 14.92	71.11 ± 5.98	11.53 ± 2.44	51.41 ± 10.66	47.89 ± 13.56	72.07 ± 3.94	8.62 ± 1.89
DeepSeek-V3.2 <sup>†</sup>	50.51 ± 8.61	38.64 ± 12.57	71.73 ± 4.62	17.33 ± 2.62	54.35 ± 8.76	47.39 ± 10.95	72.14 ± 4.56	14.09 ± 2.69
DeepSeek-V3.2-Thinking <sup>†</sup>	46.73 ± 9.68	34.62 ± 12.71	69.53 ± 4.90	16.17 ± 2.80	49.37 ± 11.12	41.98 ± 13.42	70.72 ± 4.69	12.56 ± 2.92
GLM-4.6 <sup>†</sup>	47.02 ± 8.96	34.70 ± 11.87	68.48 ± 5.00	16.73 ± 2.43	48.23 ± 11.47	40.44 ± 13.68	68.92 ± 5.43	13.31 ± 2.82
Kimi K2 Thinking <sup>†</sup>	54.05 ± 9.57	42.86 ± 12.70	70.76 ± 6.51	17.10 ± 2.89	55.26 ± 10.00	49.45 ± 10.79	70.55 ± 5.77	13.92 ± 2.60
MiniMax-M2 <sup>†</sup>	57.78 ± 11.02	46.18 ± 12.46	73.87 ± 7.05	16.41 ± 2.90	63.01 ± 10.91	56.84 ± 10.52	79.74 ± 9.28	14.81 ± 3.43
GPT-4o*	41.05 ± 9.73	29.38 ± 12.75	67.66 ± 5.20	10.83 ± 1.98	44.70 ± 9.80	34.79 ± 13.51	69.89 ± 3.66	10.84 ± 2.14
Gemini 2.5*	48.35 ± 11.45	35.58 ± 15.63	70.10 ± 9.38	17.82 ± 3.34	49.89 ± 11.10	39.03 ± 15.49	71.25 ± 4.75	14.69 ± 3.14
Qwen3-Max*	61.75 ± 9.56	53.40 ± 11.25	74.99 ± 4.57	17.73 ± 2.48	60.82 ± 7.83	57.84 ± 8.53	74.29 ± 4.19	14.66 ± 2.42
Doubao-Seed-1.6*	51.51 ± 8.85	39.80 ± 11.20	69.03 ± 5.30	18.04 ± 2.72	51.11 ± 8.90	44.37 ± 9.95	68.35 ± 4.63	14.38 ± 2.61
ERNIE-5.0-Preview*	47.09 ± 9.99	35.55 ± 11.93	67.19 ± 5.73	15.82 ± 2.69	45.05 ± 9.49	40.03 ± 11.59	67.03 ± 5.13	12.41 ± 2.67
CoT	64.72 ± 9.01	60.34 ± 7.09	76.94 ± 4.74	25.83 ± 2.97	65.61 ± 7.57	61.89 ± 7.54	78.57 ± 4.95	21.23 ± 3.55
ToT	66.53 ± 9.75	62.34 ± 6.64	78.46 ± 5.49	25.96 ± 3.14	67.58 ± 6.85	63.96 ± 6.17	79.86 ± 5.50	21.77 ± 3.13
MDAgents	68.41 ± 9.61	56.45 ± 8.53	78.81 ± 6.76	24.83 ± 2.53	66.58 ± 10.05	60.98 ± 8.92	79.95 ± 8.41	21.28 ± 2.73
MedAgents	59.85 ± 9.39	54.48 ± 7.95	72.00 ± 5.88	23.66 ± 2.33	61.51 ± 8.81	57.19 ± 7.42	74.30 ± 5.17	22.04 ± 2.33
<b>Aegle (Ours)</b>	<b>72.78 ± 10.16</b>	<b>63.02 ± 5.57</b>	<b>77.55 ± 5.41</b>	<b>25.83 ± 2.55</b>	<b>71.52 ± 8.35</b>	<b>63.92 ± 4.73</b>	<b>79.93 ± 6.90</b>	<b>24.24 ± 2.44</b>

Table 1: Documentation quality evaluation on ClinicalBench and RAPID-IPN. Models marked with † are open-source, those with ‡ are medical-domain models, and those with \* are proprietary models. Rows shaded in light gray correspond to reasoning-strategy baselines, while rows shaded in light blue correspond to medical multi-agent systems. Mean ± standard deviation is reported.

actual clinical trajectories, thereby guaranteeing real-world fidelity. Due to patient privacy regulations and institutional data governance policies, the RAPID-IPN dataset cannot be publicly released. All data were fully de-identified in accordance with local regulations prior to use, and the study protocol was reviewed and approved by the hospital’s ethics committee.

## 4.2 Experimental Setup

**Baselines and Evaluation Paradigm.** To rigorously benchmark our framework, we compare Aegle against general-purpose single-LLM baselines and advanced reasoning strategies including CoT and ToT. We further adapt specialized medical multi-agent systems, specifically MDAgents (Kim et al., 2024) and MedAgents (Tang et al., 2024), to this interactive task using DeepSeek-V3.2 as the consistent backbone to isolate architectural benefits. All models are evaluated under an LLM-as-a-judge paradigm using gpt-4o-mini, with identical scoring prompts across all conditions to ensure fair and consistent comparison. We further conducted a small-scale human evaluation to validate the reliability of the LLM-as-a-judge paradigm; details are provided in the Appendix A.

**Evaluation Metrics.** We adopt a multi-dimensional evaluation framework that assesses both the consultation process and the resulting clinical documentation. Specifically, documentation quality is evaluated along clinical reasoning

(IDEA), documentation standardization (SOAP), readability (READ), and surface-level similarity (chrF++). Consultation capability is assessed using a consultation skills rubric covering inquiry skills and humanistic care. Detailed introduction and rubrics are provided in Appendix C.

## 4.3 Documentation Quality Evaluation

The documentation quality results in Table 1 reveal a clear and consistent performance advantage for Aegle across both benchmarks. This advantage extends beyond comparisons with single-model baselines and remains evident against strong reasoning-strategy baselines (CoT and ToT) as well as existing medical multi-agent systems (MDAgents and MedAgents). Notably, the gains are most pronounced in metrics related to internal coherence and evidential grounding, reflecting a shift from surface-level summarization toward more structured clinical reasoning.

In single-model baselines, IPN often exhibit a familiar failure mode: fluent narratives that read well locally but lack global alignment between history, assessment, and plan. While CoT-style prompting partially alleviates this issue by improving local reasoning consistency, it does not explicitly constrain how evidence is accumulated and reused across sections. Aegle mitigates this limitation by enforcing an explicit separation between case feature accumulation and diagnostic synthesis. As a result, diagnostic conclusions and management plans are consistently traceable to previously documented

Model	ClinicalBench							RAPID-IPN						
	CA	QT	VER	PJ	SP	AB	Turns	CA	QT	VER	PJ	SP	AB	Turns
HuatuogPT-o1-7B <sup>†‡</sup>	4.01	4.44	4.88	4.15	4.96	4.92	21.67	4.00	4.43	4.89	4.20	4.99	4.98	19.42
Qwen3-8B-think <sup>†</sup>	4.00	4.10	4.69	4.78	4.94	4.14	7.03	4.00	4.07	4.66	4.68	4.96	4.11	6.86
Baichuan-M2-32B <sup>†‡</sup>	3.99	3.97	4.47	4.86	4.89	3.97	11.54	4.00	3.87	4.32	4.91	4.86	3.94	11.10
Lingshu-32B <sup>†‡</sup>	4.00	4.52	4.91	4.34	4.98	4.91	8.03	4.00	4.48	4.88	4.30	4.99	4.92	7.36
DeepSeek-V3.2 <sup>†</sup>	4.01	4.55	4.81	4.31	4.93	4.86	20.50	4.02	4.64	4.79	4.25	4.95	4.90	19.61
DeepSeek-V3.2-Thinking <sup>†</sup>	4.00	4.86	4.96	4.57	4.99	4.97	8.80	4.00	4.84	4.92	4.40	5.00	4.98	8.26
GLM-4.6 <sup>†</sup>	4.00	4.65	4.97	4.79	5.00	4.98	8.21	4.00	4.60	4.94	4.64	5.00	4.98	8.10
Kimi K2 Thinking <sup>†</sup>	4.02	4.88	4.93	4.58	4.99	4.97	8.96	4.01	4.85	4.87	4.38	5.00	4.97	8.14
MiniMax-M2 <sup>†</sup>	4.04	4.77	4.77	4.44	4.99	4.85	9.14	4.04	4.34	4.05	4.68	5.00	4.57	21.23
GPT-4o*	4.00	4.74	4.93	4.91	5.00	4.56	4.60	4.00	4.88	4.94	4.84	5.00	4.49	3.84
Gemini 2.5*	3.88	3.58	4.12	4.89	4.38	4.00	29.98	3.95	3.60	4.09	4.89	4.28	3.96	30.00
Qwen3-Max*	4.01	4.90	4.96	4.38	5.00	4.96	6.42	4.01	4.84	4.97	4.34	5.00	4.96	5.41
Doubao-Seed-1.6*	4.00	4.45	4.88	4.61	4.99	4.66	10.43	4.00	4.32	4.82	4.56	5.00	4.66	10.04
ERNIE-5.0-Preview*	4.00	4.55	4.92	4.59	4.97	4.39	5.06	4.00	4.36	4.84	4.47	4.97	4.36	4.51
CoT	4.00	4.83	4.63	4.05	4.99	4.85	6.40	4.00	4.78	4.48	4.28	5.00	4.86	6.66
ToT	4.00	4.86	4.65	4.00	4.99	4.89	7.37	4.00	4.84	4.56	4.15	5.00	4.91	7.91
MDAgents	4.00	4.66	4.85	4.32	4.98	4.93	6.59	4.00	4.76	4.94	4.57	5.00	4.71	6.61
MedAgents	4.00	4.58	4.90	4.55	5.00	4.80	6.96	4.00	4.62	4.92	4.57	5.00	4.71	6.61
<b>Aegle (Ours)</b>	4.02	4.95	4.94	4.21	5.00	5.00	10.16	4.03	4.96	4.93	4.19	5.00	5.00	8.84

Table 2: Consultation Capability Evaluation Results on ClinicalBench and RAPID-IPN. Metrics: CA = Conversation Arrangement; QT = Question Types; VER = Verifications; PJ = Professional Jargon; SP = Speech; AB = Amiable Behavior; Turns = number of dialogue turns. Models marked with † are open-source, those with ‡ are medical-domain models, and those with \* are proprietary models. Rows shaded in light gray correspond to reasoning-strategy baselines, while rows shaded in light blue correspond to medical multi-agent systems.

evidence, reducing omissions.

On metrics such as READ and chrF++, Aegle is comparable to the strongest reasoning baselines on both datasets, suggesting that linguistic fluency has largely saturated among modern LLMs. Consequently, further gains in clinical documentation quality depend less on wording and more on how information is structured, prioritized, and constrained.

#### 4.4 Consultation Capability Evaluation

The consultation capability results in Table 2 show that Aegle’s strengths are mainly reflected in how consultation information is elicited and validated, rather than in stylistic or expressive aspects of dialogue. Across both benchmarks, Aegle consistently demonstrates a more directed questioning strategy, in which dialogue turns are organized around resolving clinically relevant uncertainties.

When compared with single-model baselines and reasoning-strategy baselines such as CoT and ToT, Aegle exhibits a more structured pattern of information verification. Its higher VER and QT scores indicate that follow-up questions are more frequently used to confirm or refine patient-provided information, instead of extending the conversation through loosely related prompts. This pattern is characteristic of specialist-driven history taking, where each question serves a specific diag-

nostic purpose.

Conversational style and humanistic expression metrics remain close to their upper bounds for most competitive models, which limits their discriminative value. In terms of dialogue length, Aegle occupies a middle range, avoiding both very short interactions that may under-verify critical details and excessively long exchanges that dilute diagnostic focus. By maintaining this balance while achieving strong verification and question coverage, Aegle demonstrates a consultation behavior that is both efficient and clinically grounded, which is consistent across the two benchmarks.

#### 4.5 Expert Activation Efficiency

Beyond output quality, the practicality of multi-agent systems critically depends on the efficient utilization of expert resources during reasoning. Table 3 compares the average number of activated specialists across MDT-style frameworks.

Model	Architecture	Experts per Case	Experts per Round
MDAgents	Static Multi-Agent	3.702	3.702
MedAgents	Static Multi-Agent	4.968	4.968
<b>Aegle (Ours)</b>	Dynamic Virtual MDT	<b>2.416</b>	<b>1.423</b>

Table 3: Comparison of specialist activation across different multi-agent frameworks. Architecture indicates the underlying expert coordination scheme.

Static multi-agent baselines such as MDAgents

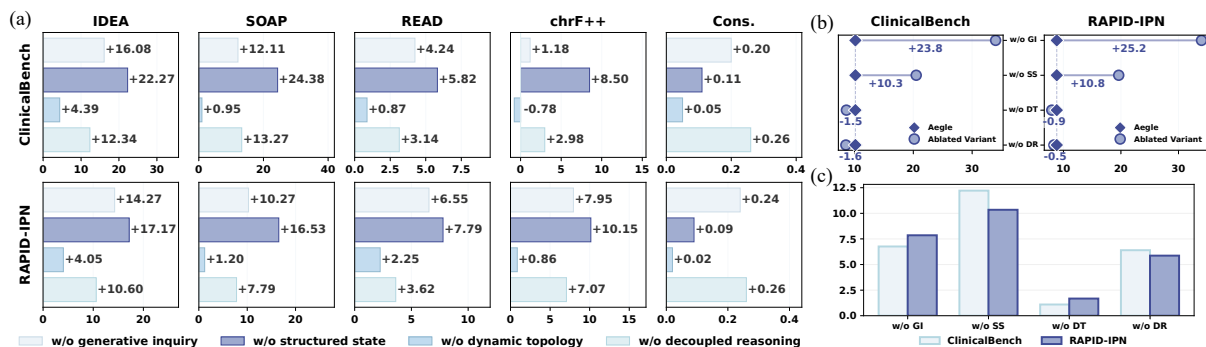


Figure 4: Ablation study results on ClinicalBench and RAPID-IPN. (a) Performance Degradation: The score drop of ablated variants compared to Aegle across five documentation quality metrics. Higher bars indicate a larger contribution of that component to the model’s performance. (b) Dialogue Efficiency: Comparison of consultation turns. Removing Generative Inquiry (w/o GI) or Structured State (w/o SS) leads to significantly longer and less efficient dialogues. (c) Average Drop: The average performance degradation across all metrics, highlighting the Structured State as the most critical component for overall quality.

464 and MedAgents activate a fixed set of experts at  
 465 every dialogue turn, resulting in identical expert  
 466 counts per case and per round. In contrast, Aegle  
 467 employs a state-aware dynamic topology that acti-  
 468 vates specialists on demand. This design substan-  
 469 tially reduces redundant expert invocation, achiev-  
 470 ing fewer activated experts per case and, more im-  
 471 portantly, per round, without sacrificing diagnos-  
 472 tic performance. The results highlight that Aegle  
 473 improves not only clinical reasoning quality but  
 474 also computational efficiency, which is essential  
 475 for real-time outpatient deployment.

#### 476 4.6 Ablation Study

477 To reveal the roles each component plays in Aegle,  
 478 we conduct comprehensive ablation experiments  
 479 and visualize the results in Figure 4.

480 As shown in Figure 4(a) and (c), among all vari-  
 481 ants, removing the structured clinical state leads to  
 482 the most severe degradation across both datasets,  
 483 particularly in IDEA and SOAP scores. This  
 484 confirms that explicitly separating case features  
 485 from diagnostic outputs is essential for maintain-  
 486 ing evidence-grounded reasoning and standardized  
 487 documentation.

488 Eliminating generative inquiry produces a dif-  
 489 ferent failure mode. In this setting, history taking  
 490 follows a fixed template based on *Bates’ Guide to*  
 491 *Physical Examination and History Taking* (Bick-  
 492 ley and Szilagyi, 2012). While surface-level docu-  
 493 mentation quality remains relatively high, reason-  
 494 ing quality deteriorates and, as illustrated in Fig-  
 495 ure 4(b), dialogue length increases dramatically.  
 496 The sharp rise in dialogue turns indicates inefficient

497 and unfocused information gathering, suggesting  
 498 that without context-aware questioning, the system  
 499 struggles to converge on a sufficiently informative  
 500 case representation.

501 By comparison, removing dynamic topology or  
 502 decoupled reasoning results in more moderate but  
 503 systematic performance drops (Figure 4(a)). With-  
 504 out dynamic specialist activation, the system loses  
 505 adaptability to case-specific ambiguity, while re-  
 506 moving decoupled reasoning reduces hypothesis  
 507 diversity and increases the risk of premature con-  
 508 vergence. Taken together, these results suggest  
 509 that Aegle’s performance gains do not arise from  
 510 any single mechanism, but from the coordinated  
 511 interaction of structured state representation, active  
 512 inquiry, adaptive expert selection, and independent  
 513 specialist reasoning.

#### 514 5 Conclusion

515 In this paper, we presented Aegle, a virtualized  
 516 MDT framework that elevates the quality of outpa-  
 517 tient consultation by enabling synchronous, multi-  
 518 perspective reasoning within the intake workflow.  
 519 By enforcing a structural separation between evi-  
 520 dence acquisition and diagnostic synthesis within  
 521 a dynamic multi-agent topology, Aegle effectively  
 522 mitigates the cognitive biases and premature clo-  
 523 sure inherent in single-view models. Extensive  
 524 evaluations on ClinicalBench and RAPID-IPN  
 525 dataset demonstrate that Aegle consistently out-  
 526 performs state-of-the-art baselines in both docu-  
 527 mentation quality and consultation capability, es-  
 528 tablishing a robust and scalable paradigm for next-  
 529 generation clinical decision support systems.

## 530 Limitations

531 Despite its strong empirical performance, Aegle  
532 has several limitations that warrant careful consid-  
533 eration. First, the multi-agent paradigm inevitably  
534 introduces additional inference overhead. Dynamic  
535 routing, parallel specialist execution, and structured  
536 aggregation increase end-to-end latency compared  
537 with single-model generation, which may prolong  
538 user waiting time in real-time outpatient settings  
539 where responsiveness is critical. Second, maintain-  
540 ing a state-aware collaboration grounded in a con-  
541 tinuously evolving SOAP record leads to longer  
542 effective contexts. As the dialogue progresses,  
543 the accumulated structured state and intermediate  
544 agent outputs expand the context window, increas-  
545 ing token consumption and computational cost, and  
546 potentially constraining deployment under strict  
547 resource budgets. Third, while fully decoupled par-  
548 allel reasoning is central to preserving hypothesis  
549 diversity and mitigating premature convergence,  
550 it can also yield redundant or overlapping recom-  
551 mendations across specialists, particularly in cases  
552 with shared clinical priors or insufficiently differ-  
553 entiated task instructions. Such repetition not only  
554 wastes computation but may also complicate aggre-  
555 gation by diluting genuinely novel signals. Future  
556 work should therefore explore mechanisms that bet-  
557 ter balance cognitive bias mitigation and content  
558 redundancy, such as diversity-aware expert prompt-  
559 ing, redundancy-penalized aggregation, or adaptive  
560 expert selection strategies.

## 561 Acknowledgments

562 The acknowledgments are omitted for the review  
563 process.

## 564 References

565 Elizabeth A. Baker, Cynthia H. Ledford, Louis Fogg,  
566 David P. Way, and Yoon Soo Park. 2015. *The IDEA*  
567 *assessment tool: Assessing the reporting, diagnostic*  
568 *reasoning, and decision-making skills demonstrated*  
569 *in medical students' hospital admission notes. Teaching*  
570 *and Learning in Medicine*, 27(2):163–173.

571 Lynn Bickley and Peter G. Szilagyi. 2012. *Bates' guide*  
572 *to physical examination and history-taking*. Lippin-  
573 cott Williams & Wilkins.

574 Katherine C. Brooks, Katie E. Raffel, David Chia, Ab-  
575 hishek Karwa, Colin C. Hubbard, Andrew D. Auer-  
576 bach, and Sumant R. Ranji. 2024. *Stigmatizing lan-*  
577 *guage, patient demographics, and errors in the diag-*  
578 *nostic process. JAMA Internal Medicine*, 184(6):704–  
579 706.

Xi Chen, Huahui Yi, Mingke You, WeiZhi Liu, Li Wang,  
Hairui Li, Xue Zhang, Yingman Guo, Lei Fan, Gang  
Chen, Qicheng Lao, Weili Fu, Kang Li, and Jian Li.  
2025. *Enhancing diagnostic capability with multi-*  
*agents conversational large language models. npj*  
*Digital Medicine*, 8(1):159.

Pat Croskerry. 2013. *From mindless to mindful practice:*  
*Cognitive bias and clinical decision making. New*  
*England Journal of Medicine*, 368(26):2445–2448.

DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin,  
Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao  
Wu, Bowei Zhang, Chaofan Lin, Chen Dong,  
Chengda Lu, Chenggang Zhao, Chengqi Deng, Chen-  
hao Xu, Chong Ruan, Damai Dai, Daya Guo, De-  
jian Yang, and 193 others. 2025. *DeepSeek-V3.2:*  
*Pushing the frontier of open large language models.*  
*Preprint*, arXiv:2512.02556.

Lei Gu, Yinghao Zhu, Haoran Sang, Zixiang Wang,  
Dehao Sui, Wen Tang, Ewen Harrison, Junyi Gao,  
Lequan Yu, and Liantao Ma. 2025. *MedAgentAu-*  
*dit: Diagnosing and quantifying collaborative failure*  
*modes in medical multi-agent systems. Preprint,*  
arXiv:2510.10185.

Xudong Han, Xianglun Gao, Xiaoyi Qu, and Zhenyu  
Yu. 2025. *Multi-agent medical decision consensus*  
*matrix system: An intelligent collaborative frame-*  
*work for oncology MDT consultations. Preprint,*  
arXiv:2512.14321.

Health Human Resources Development Center, Na-  
tional Health Commission of China. 2024. *Standard*  
*scheme for the final clinical practice ability assess-*  
*ment of standardized resident training.*

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 for medical decision-making. In Advances*  
*in Neural Information Processing Systems. Oral pre-*  
*sentation.*

Kundan Krishna, Sapan Khosla, Jeffrey Bigham, and  
Zachary C. Lipton. 2021. *Generating SOAP notes*  
*from doctor-patient conversations using modular*  
*summarization techniques. In Proceedings of the*  
*59th Annual Meeting of the Association for Compu-*  
*tational Linguistics and the 11th International Joint*  
*Conference on Natural Language Processing (Vol-*  
*ume 1: Long Papers)*, pages 4958–4972, Online. As-  
sociation for Computational Linguistics.

Maya Kruse, Shiyue Hu, Nicholas Derby, Yifu Wu,  
Samantha Stonbraker, Bingsheng Yao, Dakuo Wang,  
Elizabeth M. Goldberg, and Yanjun Gao. 2025.  
*Large language models with temporal reasoning for*  
*longitudinal clinical summarization and prediction.*  
*In Findings of the Association for Computational Lin-*  
*guistics: EMNLP 2025*, pages 20715–20735, Suzhou,  
China. Association for Computational Linguistics.

636	Pengcheng Qiu, Chaoyi Wu, Junwei Liu, Qiaoyu Zheng,	Yijie Wang, Yining Chen, and Jifang Sheng. 2024. <i>As-</i>	691
637	Yusheng Liao, Haowen Wang, Yun Yue, Qianrui Fan,	sessing ChatGPT as a medical consultation assis-	692
638	Shuai Zhen, Jian Wang, Jinjie Gu, Yanfeng Wang,	tant for chronic hepatitis B: Cross-language study	693
639	Ya Zhang, and Weidi Xie. 2025. <i>Evolving diagnostic</i>	of english and chinese. <i>JMIR Medical Informatics</i> ,	694
640	<i>agents in a virtual clinical environment. Preprint,</i>	12:e56426.	695
641	arXiv:2510.24654.		
642	Zhiyao Ren, Yibing Zhan, Baosheng Yu, Liang Ding,	Weixiang Yan, Haitian Liu, Tengxiao Wu, Qian Chen,	696
643	Pingbo Xu, and Dacheng Tao. 2025. <i>Healthcare</i>	Wen Wang, Haoyuan Chai, and Jiayi Wang. 2025.	697
644	<i>agent: Eliciting the power of large language models</i>	<i>ClinicalLab: Aligning agents for multi-departmental</i>	698
645	<i>for medical consultation. npj Artificial Intelligence,</i>	<i>clinical diagnostics in the real world. In The Thirty-</i>	699
646	1(1):24.	<i>ninth Annual Conference on Neural Information Pro-</i>	700
		<i>cessing Systems.</i>	701
647	Sara Salehi, Yashbir Singh, Parnian Habibi, and	Jacqueline G. You, Reema H. Dbouk, Adam Land-	702
648	Bradley J. Erickson. 2025. <i>Beyond single systems:</i>	man, David Y. Ting, Sayon Dutta, Julie C. Wang,	703
649	<i>How multi-agent AI is reshaping ethics in radiology.</i>	Amanda J. Centi, Molly Macfarlane, Eran Bechor,	704
650	<i>Bioengineering</i> , 12(10):1100.	Jonathan Letourneau, Gabrielle Choo-Kang, Es-	705
651	Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres,	ther H. Kim, Cordula Magee, Brian J. Lang, Laura	706
652	Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin	Angelo, Jackson Olin, Michelle Frits, Christine Ian-	707
653	Clark, Stephen R. Pfohl, Heather Cole-Lewis, Dar-	naccone, Angela Rui, and 7 others. 2025. <i>Ambient</i>	708
654	lene Neal, Qazi Mamunur Rashid, Mike Schaeker-	<i>documentation technology in clinician experience of</i>	709
655	mann, Amy Wang, Dev Dash, Jonathan H. Chen,	<i>documentation burden and burnout. JAMA Network</i>	710
656	Nigam H. Shah, Sami Lachgar, Philip Andrew Mans-	<i>Open</i> , 8(8):e2528056.	711
657	field, and 16 others. 2025. <i>Toward expert-level medi-</i>		
658	<i>cal question answering with large language models.</i>	Sihang Zeng, Yujian Fu, Sitong Zhou, Zixuan Yu, Lu-	712
659	<i>Nature Medicine</i> , 31(3):943–950.	cas Jing Liu, Jun Wen, Matthew Thompson, Ruth	713
660	Barbara Starfield, Leiyu Shi, and James Macinko. 2005.	Etzioni, and Meliha Yetisgen. 2025. <i>Traj-CoA: Pa-</i>	714
661	<i>Contribution of primary care to health systems and</i>	<i>tient trajectory modeling via chain-of-agents for lung</i>	715
662	<i>health. The Milbank Quarterly</i> , 83(3):457–502.	<i>cancer risk prediction. In NeurIPS 2025 Workshop</i>	716
		<i>on GenAI4Health.</i>	717
663	Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming	Juexiao Zhou, Haoyang Li, Siyuan Chen, Zhangtianyi	718
664	Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and	Chen, Zhongyi Han, and Xin Gao. 2025a. <i>Large</i>	719
665	Mark Gerstein. 2024. <i>MedAgents: Large language</i>	<i>language models in biomedicine and healthcare. npj</i>	720
666	<i>models as collaborators for zero-shot medical rea-</i>	<i>Artificial Intelligence</i> , 1(1):44.	721
667	<i>soning. In Findings of the Association for Com-</i>		
668	<i>putational Linguistics: ACL 2024</i> , pages 599–621,	Zhanke Zhou, Xiao Feng, Zhaocheng Zhu, Jiangchao	722
669	Bangkok, Thailand. Association for Computational	Yao, Sanmi Koyejo, and Bo Han. 2025b. <i>From pas-</i>	723
670	Linguistics.	<i>sive to active reasoning: Can large language models</i>	724
		<i>ask the right questions under incomplete informa-</i>	725
671	Cath Taylor, Alastair J. Munro, Rob Glynne-Jones,	<i>tion? In Proceedings of the Forty-second Interna-</i>	726
672	Clive Griffith, Paul Trevatt, Michael Richards, and	<i>tional Conference on Machine Learning.</i>	727
673	Amanda J. Ramirez. 2010. <i>Multidisciplinary team</i>		
674	<i>working in cancer: What is the evidence? BMJ</i> ,		
675	340:c951.		
676	Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab,		
677	Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li,		
678	Mohamed Amin, Yong Cheng, Elahe Vedadi, Nenad		
679	Tomašev, Shekoofeh Azizi, Karan Singhal, Le Hou,		
680	Albert Webson, Kavita Kulkarni, S. Sara Mahdavi,		
681	Christopher Semturs, and 7 others. 2025. <i>Towards</i>		
682	<i>conversational diagnostic artificial intelligence. Na-</i>		
683	<i>ture</i> , 642(8067):442–450.		
684	Chen Wang, Shuang Li, Ning Lin, Xin Zhang, Yu Han,		
685	Xia Wang, Dong Liu, Xue Tan, Di Pu, Kun Li, Gang		
686	Qian, and Rui Yin. 2025. <i>Application of large lan-</i>		
687	<i>guage models in medical training evaluation: Us-</i>		
688	<i>ing ChatGPT as a standardized patient—multimetric</i>		
689	<i>assessment. Journal of Medical Internet Research</i> ,		
690	27:e59435.		

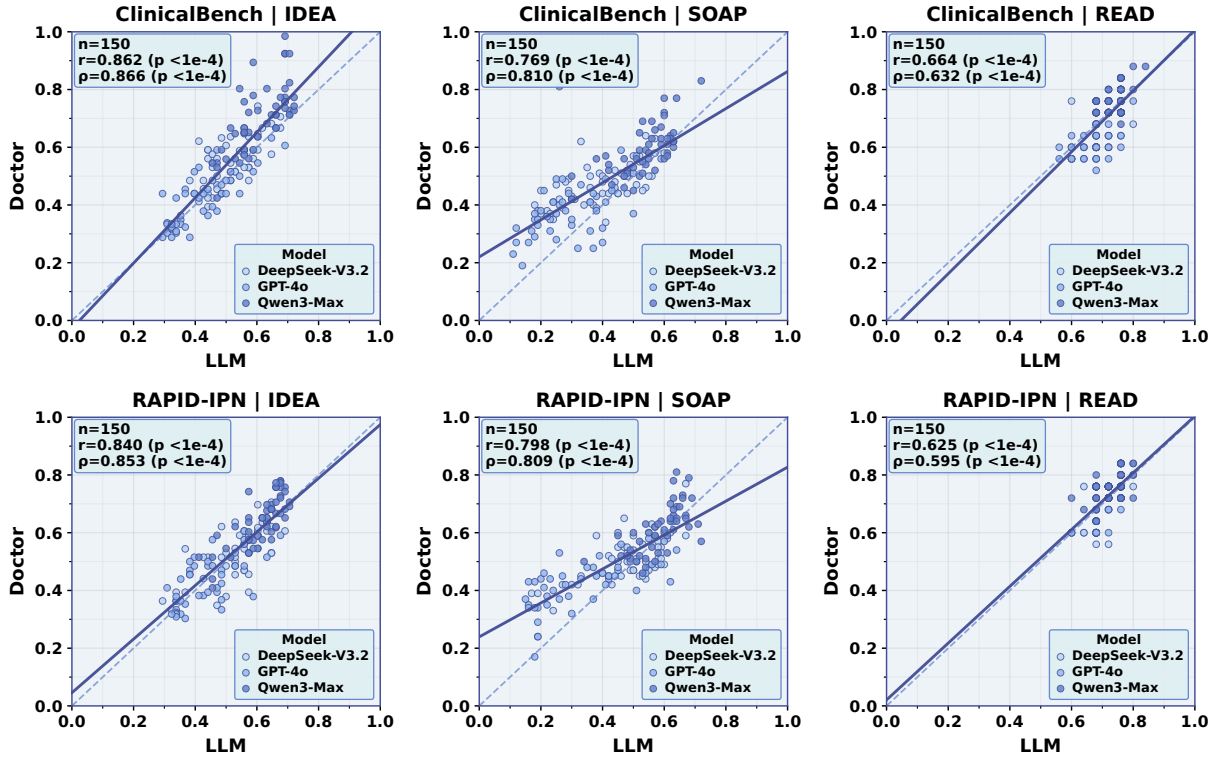


Figure 5: Correlation between LLM-as-a-judge scores and physician ratings on ClinicalBench and RAPID-IPN. Each point represents one evaluated IPN instance ( $n = 150$  per dataset). Pearson ( $r$ ) and Spearman ( $\rho$ ) correlation coefficients are reported for IDEA, SOAP, and READ metrics.

## A Analysis of the reliability of LLM-as-a-judge

To assess the reliability of the LLM-as-a-judge evaluation paradigm used throughout our experiments, we conducted a targeted human evaluation study and compared physician ratings with the scores produced by the judge model.

For each dataset (ClinicalBench and RAPID-IPN), we randomly sampled 50 cases and selected three representative models for comparison (DeepSeek-V3.2, GPT-4o, and Qwen3-Max), resulting in a total of 150 evaluated instances per dataset. Licensed physicians independently scored the generated IPNs using the same evaluation rubrics as those employed in the automatic assessment, covering IDEA, SOAP, and READ dimensions.

Figure 5 presents the correlation analysis between physician scores and LLM-as-a-judge scores.

Across both datasets and all evaluated metrics, we observe statistically significant positive correlations. On ClinicalBench, Pearson correlation coefficients range from 0.664 (READ) to 0.862 (IDEA), with corresponding Spearman rank correlations

ranging from 0.632 to 0.866. On RAPID-IPN, Pearson correlations range from 0.625 (READ) to 0.840 (IDEA), with Spearman correlations ranging from 0.595 to 0.853. All correlations are significant with  $p < 10^{-4}$ .

These results indicate that the LLM-as-a-judge scores are broadly consistent with physician judgments across different datasets and evaluation dimensions. While minor discrepancies remain, the overall alignment supports the validity of using LLM-based evaluation as a scalable proxy for large-scale comparative experiments.

## B Prompt Design

We present the prompt design for Standardized Patient, Orchestrator, Specialist, and Aggregator in this section.

## Standardized Patient

You are a professionally trained standardized patient actor, simulating a realistic medical consultation with a physician.

Below is your complete case information:

<Case Information>

{case\_information}

</Case Information>

### **\*\*Role-Playing Principles\*\***

1. **\*\*Based on the information\*\***: All of your responses must be strictly based on the case text above.
2. **\*\*No fabrication\*\***: Do not imagine, extend, or invent any medical history or details that are not provided in the case. If the doctor asks about information that is completely absent from the materials, respond naturally with phrases such as “I’m not sure,” “I didn’t notice,” or “I don’t have that issue.”
3. **\*\*Non-professional stance\*\***: You are an ordinary patient and should not provide medical explanations, professional interpretations, or diagnostic suggestions.
4. **\*\*Natural expression\*\***: Use conversational, natural, emotionally realistic first-person language (“I...”). Do not list information like a robot.

### **\*\*Expression Requirements\*\***

- Respond as a real patient speaking to a doctor, avoid mechanical answers, lists, or summaries.
- When describing experiences, symptoms, and feelings, focus on subjective experience (e.g., severity, duration, impact, emotions).
- If professional terms are involved (such as test results mentioned in the case), paraphrase them naturally, or only read them out in detail when the doctor specifically asks for the report.

### **\*\*Exceptions\*\***

- For objective information such as test results, laboratory data, or imaging findings, you may report them truthfully according to the case when the doctor asks.
- If the doctor presses for specific details that do not exist in the case materials (e.g., exact numerical values or redacted information), clearly state that you cannot provide them.

You will now answer the doctor’s questions in the role of the patient.

## Orchestrator

You are a Medical MDT Orchestrator.

**Current phase: {current\_phase\_description}**

Phase goal: {phase\_goal}

Your responsibilities:

- Efficient decision-making**: For simple confirmation questions, routine information gathering, or single-dimension follow-up inquiries, you should generate the question directly without activating specialist physicians.
- For complex situations, dynamically dispatch specialist Agents (Specialists) and issue them instructions that **align with the current phase goal**.

**Decision logic**:

- Simple cases**: When only specific numerical values need to be supplemented, past medical history needs to be confirmed, or simple details need to be clarified → **keep `active\_specialists` empty** and place the question directly into `suggested\_question`.
- Complex cases**: When differential diagnosis, multi-system symptoms, or in-depth analysis is required → activate the relevant Specialists.

**Strictly output in the specified JSON format**

**You do not make medical diagnoses; you are responsible only for coordination and orchestration.**

### historytaking

current\_phase\_description:

Phase 1: History Taking

phase\_goal:

Efficiently collect the patient's case features, fill in information gaps, and clarify symptom details. Focus on documenting the patient's actual clinical information: basic information, history of present illness, past medical history, physical examination, and ancillary tests. Note: this is an information-gathering and documentation phase, not a phase for formulating investigation plans.

### diagnostic\_synthesis

current\_phase\_description:

Phase 2: Diagnosis & Plan

phase\_goal:

Based on the established case features, organize discussions among specialists to provide a preliminary diagnosis, differential diagnostic considerations, and a diagnostic and treatment plan.

## Specialist

You are a specialist physician focusing on the field of {spec\_id}. As a member of a multidisciplinary team (MDT), your core responsibility is to assist in refining the case information and to provide precise diagnostic and treatment recommendations from the professional perspective of {spec\_id}.

Tasks of the current phase:  
{phase\_instructions}

Medical record documentation guidance:  
{soap\_guidance}

Your workflow:

1. Analyze the input information and determine which SOAP field each piece of information should be written into.
2. Strictly output in the JSON format defined by `SpecialistOutput`, and in `draft\_modifications` clearly specify the exact SOAP field to which each item of information belongs, for example:
  - "Write 'chest pain for 2 hours' into `history\_of\_present\_illness`"
  - "Write 'blood pressure 150/90 mmHg' into `physical\_examination`"
3. {next\_question\_instruction}

### historytaking

#### phase\_instructions:

From the professional perspective of {spec\_id}, carefully review the current case features. Focus on clinical information that is directly relevant to your specialty, including basic patient information, history of present illness, past medical history, physical examination findings, and ancillary test results.

If the existing information is not sufficient to fully and accurately document the patient's condition, raise specific and concrete interview questions to supplement the missing details. If the available information is already adequate, clearly state that there are no additional questions.

At this stage, your role is strictly limited to recording and refining clinical information; do not propose diagnostic tests or treatment plans.

#### next\_question\_instruction:

Fill in the questions that need to be raised in 'suggested\_questions' to advance the consultation. If there are no questions, you should state the end.

### diagnostic\_synthesis

#### phase\_instructions:

Based on the finalized case features, provide your professional opinion from the perspective of {spec\_id}.

In the suggestions\_for\_draft\_revision section, structure your output strictly as follows:

- Preliminary diagnosis
- Diagnostic discussion (including diagnosis basis and differential diagnosis)
- Treatment plan, covering further investigations, pharmacologic therapy, non-pharmacologic interventions, and follow-up requirements

#### next\_question\_instruction:

Fill in 'N/A' in the 'suggested\_questions' field and no longer ask questions to the patient.

## historytaking

### soap\_guidance:

#### Basic Information (basic\_information):

- Patient demographics and chief complaint
- Includes: age, sex, occupation, visit date, main symptoms
- Keep it concise and focused on the main problem

#### History of Present Illness (history\_of\_present\_illness):

- Onset and course of the current illness
- Includes: how it started, symptom characteristics, progression, prior evaluation and treatment, and response
- Describe in chronological order with clear logic

#### Past Medical History (past\_medical\_history):

- Patient's previous health history
- Includes: past diseases, surgeries, trauma, allergies, personal and family history
- Focus on information related to the current condition

#### Physical Examination (physical\_examination):

- Objective findings from the physician's examination
- Includes: vital signs and system-based exam findings
- Describe objectively; avoid subjective judgments

#### Auxiliary Examination (auxiliary\_examination):

- Laboratory and imaging results
- Includes: test values and examination dates
- Report objective data accurately

## diagnostic\_synthesis

### soap\_guidance:

#### DIAGNOSIS & PLAN:

#### Preliminary Diagnosis (preliminary\_diagnosis):

- List diagnoses line by line in standard clinical format
- Include major and relevant secondary diagnoses
- Order by importance or timeline

#### Diagnostic Discussion (diagnosis\_discussion):

- Explain diagnostic reasoning and differential diagnosis
- Summarize key evidence: patient profile, symptoms, exam findings, and test results
- Clearly mention:
  - Risk factors or health problems (if unknown, state "unclear")
  - Complications or related conditions
  - Treatment adherence
  - Available family or social support
- Do not invent tests or list unmasked negative findings

#### Treatment Plan (treatment\_plan):

##### 1. Further evaluation:

- Planned tests, purpose, and timing, with guideline-based rationale
- If surgery is indicated, include indication and approach

##### 2. Medications:

- Drug classes with indications, contraindications, and monitoring points
- Avoid specific dosages

##### 3. Non-drug management:

- Hydration goals, diet, lifestyle advice, pain management, and warning signs

##### 4. Follow-up:

- Follow-up timeframe, repeat tests, and visit format (in-person or remote)

## Aggregator

You are a recorder and decision-maker in a medical MDT.

**Current phase: {phase\_name}**

As a professional physician, you must organize and document patient information strictly according to the standard SOAP format for the initial medical record:

{soap\_guidance}

Your task:

{aggregator\_task\_description}

**Work requirements:**

1. **Information integration**: Carefully review the input from specialist physicians (if any). **If 'specialist\_outputs' is empty, extract information directly from the patient's latest response and the coordinator's suggestions, and update the SOAP record accordingly.**
2. **Empathy**: You are the doctor directly communicating with the patient; your tone should be warm, professional, and patient.
3. **Logical consistency**: Maintain a complete and coherent medical record. If the patient has clearly stated that certain information cannot be provided, document this truthfully and do not ask again.

**Strictly output in the specified JSON format**

### historytaking

soap\_guidance:

(the same as specialist)

aggregator\_task\_description:

Summarize the question suggestions from specialist physicians and ask the patient accordingly.

When the patient mentions key symptoms, make sure to confirm them.

The patient is a layperson, so avoid overly technical or uncommon medical terms.

### diagnostic\_synthesis

soap\_guidance:

(the same as specialist)

aggregator\_task\_description:

Integrate all specialist opinions to generate preliminary diagnosis, diagnostic discussion, and treatment plan.

## C Evaluation Metrics

We adopt a multi-dimensional evaluation framework that assesses both the consultation process and the resulting clinical documentation. All models are evaluated under an LLM-as-a-judge paradigm using gpt-4o-mini. Identical scoring prompts and rubric definitions are applied across all settings to ensure consistency and fair comparison.

**Documentation Quality.** We evaluate the quality of generated SOAP notes from complementary perspectives covering clinical reasoning, documentation standardization, readability, and reference similarity. Specifically, we use the following metrics.

- **IDEA Score (Baker et al., 2015).** IDEA evaluates the completeness and coherence of clinical reasoning by examining alignment among history taking, physical examination, diagnosis and differential, and care planning. Higher scores require detailed and well-organized HPI, complete and diagnostically relevant physical examinations, diagnoses supported by objective evidence with clear reasoning and ranked differentials, and comprehensive, appropriate care plans. Internal inconsistencies across sections are explicitly penalized. The detailed rubric is provided in Table 4.
  - **SOAP Score (Health Human Resources Development Center, National Health Commission of China, 2024).** SOAP Score measures adherence to standardized SOAP documentation practices. The rubric evaluates completeness and accuracy within each section. The *Subjective* component emphasizes structured problem descriptions, concise chief complaints, detailed symptom characterization, prior evaluations and treatments, and relevant medical, family, and social histories. The *Objective* component focuses on complete physical examinations and accurate reporting of laboratory and ancillary tests. The *Assessment* component rewards clear diagnoses justified by clinical evidence, analysis of risk factors and comorbidities, and evaluation of adherence and family resources. The *Plan* component emphasizes guideline-consistent diagnostic and management plans, detailed treatment strategies, non-pharmacologic interventions, and explicit follow-up requirements.
- Predefined deduction codes are used to annotate common documentation errors for fine-grained analysis. The detailed rubric and deduction codes are shown in Tables 5 and 6.
- **READ Score (Wang et al., 2024).** READ Score assesses presentation quality and clinical usability, focusing on structural completeness, logical coherence, terminology accuracy, information redundancy, and salience of key findings. It reflects how easily a note can be read, understood, and safely used in practice. The detailed rubric is presented in Table 7.
  - **chrF++.** chrF++ measures surface-level similarity between generated notes and gold-standard documentation using character  $n$ -gram overlap, providing a complementary lexical similarity signal.

**Consultation Capability.** We operationalize consultation capability using the standardized patient (SP) consultation skills grading and scoring criteria from (Wang et al., 2025). The rubric adopts a five-tier scale (5 = best) and covers two domains: *inquiry skills* and *humanistic care*. For inquiry skills, we score (i) conversation arrangement, focusing on whether the consultation has a clear opening, structured middle, and explicit closing with an orderly question flow; (ii) question types, emphasizing appropriate and balanced use of open-ended and closed-ended questions while avoiding sequential leading questions; (iii) verifications, assessing whether the clinician adequately verifies key information and cross-checks details through follow-up and reference; and (iv) professional jargon, rewarding clear patient-friendly explanations with minimal unnecessary medical terminology. For humanistic care, we score (v) speech, evaluating whether tone and pace are comfortable and appropriate, and (vi) amiable behavior, assessing whether the clinician provides empathetic responses and comfort when appropriate. We score each item by matching dialogue behaviors to tier descriptors, then aggregate item scores into an overall consultation capability score.

Sec.	Item	Definition	Max	0	1	2	3	4	
<b>1. Present Illness &amp; Comprehensive History</b>									
1	1.1 Detailed HPI	Complaint: location, quality, severity, duration, onset, radiation, aggravating/relieving factors.	4	Incorrect/illogical; key elements missing	Some elements present	Most elements present	All elements present	Concise, organized, diagnostically salient	
1	1.2 Prior Dx/Tx Course	Prior evaluations/treatments: time, location, tests, meds, interventions; full if first visit.	4	Incorrect/illogical description	Some elements present	Most elements present	All elements present	Concise, structured, informative	
1	1.3 Descriptive HPI Language	Medical descriptors (e.g., acute/chronic, sharp/dull, constant/intermittent).	3	Inappropriate/none	Minimal but appropriate	Frequent, appropriate	Consistent, accurate, concise	–	
1	1.4 Chronological Organization	Temporal ordering and coherent illness narrative.	3	Temporal contradictions	Disorganized timeline	Mostly coherent; minor gaps	Clear and consistent chronology	–	
1	1.5 Contextualized HPI	Integrates relevant PMH, family/social history, and associated symptoms.	4	No/incorrect integration	Partial integration	Comprehensive integration	Clear, accurate, concise	Key info prioritized	
1	1.6 Comprehensive History	PMH, family history, social history, review of systems.	3	Major errors/omissions	Significant missing content	Mostly complete	Thorough and complete	–	
1	1.7 Calibration	Internal consistency across the note.	–	<b>Deduction:</b> -2 points per internal inconsistency.					
<b>2. Physical Examination</b>									
2	2.1 Complete Physical Exam	Comprehensive documentation of physical examination.	4	Errors/not addressed	Major components missing	Mostly complete; minor omissions	Complete exam	Well-organized; professional terms	
2	2.2 Key Physical Findings	Highlights diagnostically relevant positive/negative findings.	3	Missing/incorrect emphasis	Partial emphasis	Comprehensive emphasis	Prioritized by relevance	–	
<b>3. Diagnosis &amp; Differential</b>									
3	3.1 Diagnostic Completeness	Primary, secondary, and additional diagnoses.	4	Primary missing/incorrect	Primary only	Primary + some secondary	Primary + all secondary	Includes additional diagnoses	
3	3.2 Objective Evidence	Evidence from history, exam, and investigations.	4	Missing/incorrect evidence	One domain	Two domains	All relevant domains	Also supports other dx	
3	3.3 Diagnostic Reasoning	Reasoning for primary diagnosis.	3	None/incorrect	Basic, partial explanation	Rigorous explanation	Links features to presentation	–	
3	3.4 Explanatory Summary	Links diagnoses, risks, and complications.	3	None/incorrect	Partial analysis	All associations discussed	Clear, logical synthesis	–	
3	3.5 Differentials	≥ 3 relevant differentials ranked by likelihood.	3	Irrelevant	< 3 or missing key alts	All key alts included	Ordered by likelihood	–	
3	3.6 Differential Reasoning	Inclusion/exclusion rationale; confounders.	3	None/incorrect	Exclusion only	Adequate exclusion only	Inclusion + exclusion; confounders	–	
3	3.7 Overall Impression	Professionalism, clarity, logical rigor.	4	Poor professionalism/logic	Adequate professionalism	Clear and consistent	Strong professional quality	Concise, polished, structured	
<b>4. Plan</b>									
4	4.1 Plan Completeness	Investigations, treatment, lifestyle, follow-up.	4	Missing/incorrect	Single aspect	Multi-dimensional	Dynamic assessment/prognosis	Concise, rigorous	
4	4.2 Plan Appropriateness	Evidence/reasoning supports key decisions.	3	Inappropriate/incorrect	Vague/unsupported	Generally appropriate	Clear; strong evidence	–	
<b>5. Overall Competency</b>									
5	5.1 Presentation Skill	Quality of written presentation.	3	–	Basic: partial	Good: most	Excellent: nearly all	–	
5	5.2 Reasoning Skill	Quality of diagnostic reasoning.	3	–	Basic reasoning	Relevant comparison	Comprehensive, rigorous	–	
5	5.3 Decision Skill	Quality of decisions in the plan.	3	–	List actions only	Partial reasoning	Evidence-based; patient-centered	–	

Table 4: IDEA scoring rubric for clinical note evaluation.

Item	Max	Rubric
<b>S: Subjective</b>		
S-1. Format	5	Each major health problem is described separately with clear categorization (e.g., somatic vs. psychological). Fully described problems score 5. If categories are mostly clear but some descriptions are brief, deduct 2–3. If key visit information or diagnostic/treatment details are omitted, categories are confused, or descriptions are fragmented, deduct 4–5.
S-2.1 Chief complaint	2	Concise and accurate summary of the primary discomfort and duration (2). If it is generally clear but not concise, or the duration is vague, deduct 1. If unclear or fails to reflect the main problem, score 0.
S-2.2 Symptoms and clinical course	5	Detailed symptom characteristics (location, quality, severity), frequency, aggravating/relieving factors, and illness trajectory (5). If key information is partially missing, deduct 2–3. If only symptoms are briefly mentioned without describing progression, score 0–2.
S-2.3 Prior evaluation and treatment	3	Prior care is documented, including facility, tests (name/time), diagnoses, medications (name/dose/duration), and response (3). If brief, deduct 1–2. If absent, score 0.
S-2.4 Relevant medical history	3	Comprehensive and accurate past history, including prior diseases, surgeries/trauma, and allergies (3). If 1–2 important elements are missing, deduct 1–2. If largely absent, score 0.
S-2.5 Family history	2	Clear documentation of heritable diseases in family members (2). If the key hereditary history is missing, deduct 1. If absent, score 0.
S-2.6 Lifestyle, psychological, and social factors	5	Comprehensive description of diet, sleep, exercise, smoking/alcohol use, mental status, work stress, family relationships, and financial situation (5). If 1–2 key elements are missing, deduct 2–3. If only briefly listed, score 0–2.
<b>O: Objective</b>		
O-1. Physical examination	8	Vital signs and system examinations are accurately and completely documented; abnormal findings are described in detail (8). If 1–2 items are missing or inaccurate, deduct 2–4. If largely missing or incorrect, score 0–3.
O-2. Laboratory and ancillary tests	5	Test items, timing, and results (values or abnormal flags) are complete and accurate (5). If 1–2 results are missing or incorrectly transcribed, deduct 2–3. If absent or disorganized, score 0–2.
O-3. Psychological tests/other assessments	2	If performed, psychological tests are documented with name and results (score/conclusion) (2). If incomplete, deduct 1. If not performed or not documented, score 0.
<b>A: Assessment</b>		
A-1. Preliminary diagnoses	4	Primary diagnosis and comorbid/secondary diagnoses are clear and complete (4). Primary diagnosis correct (2). Some secondary diagnoses are missing (1). Secondary diagnoses complete (1).
A-2.1 Diagnostic evidence	4	Diagnoses are justified using symptoms, signs, and test results with standard terminology (4). If evidence is insufficient or terminology is non-standard, deduct 1–2. If the diagnosis is incorrect or unsupported, score 0–1.
A-2.2 Risk factors and health problems	10	Disease-related risk factors and other potential health problems are comprehensively identified and their relationships analyzed (10). If 1–2 items are missing or the analysis is weak, deduct 3–5. If only listed without analysis, score 0–4.
A-2.3 Complications and comorbidities	4	Existing or potential complications and comorbidities are accurately identified and interactions analyzed (4). If important conditions are missed, deduct 2. If not analyzed, score 0–2.
A-2.4 Adherence/compliance	2	Treatment adherence is assessed based on clinical course with reasonable analysis (2). If brief, deduct 1. If incorrect or absent, score 0.
A-2.5 Family resources	1	Available family support resources (human, financial, informational) are clearly described (1). If vague, deduct 0.5. If absent, score 0.
<b>P: Plan</b>		
P-1. Further diagnostic and management plan	6	Guideline-consistent plans specify required tests, follow-up timing, and necessary consultations (6). If 1–2 key elements are missing or timing is unclear, deduct 2–3. If disorganized or generic, score 0–3.
P-2.1 Treatment plan (medications/surgery)	10	Medication or surgical plans match diagnoses, with complete details and cited guideline sources and evidence levels (10). If key information is missing, deduct 3–5. If unreasonable or largely missing, score 0–4.
P-2.2 Non-pharmacologic treatment	15	Behavioral, dietary, and exercise interventions are specific and feasible, with precautions and cited evidence (15). If overly general, deduct 5–8. If empty or vague, score 0–6.
P-3. Follow-up requirements	4	Follow-up timing and content (re-evaluation items and assessment focus) are clearly specified (4). If either timing or content is missing, deduct 2. If absent, score 0.

Table 5: SOAP scoring rubric for clinical note evaluation.

Code	Meaning	Code	Meaning
A1	Misuse of terminology	A2	Vague expression
B1	Missing important positive findings	B2	Redundant minor positive findings
C1	Negative stated as positive	C2	Positive stated as negative
C3	Missing important negative findings	D1	Irrelevant information
E1	Missing time information	E2	Vague time information
F1	Incorrect order/sequence	F2	Incorrect time value
G1	Incomplete citation of external records	G2	Incorrect paraphrase of external records
G3	Non-standard citation format	H	Compound error
I	Logical inconsistency/disorder	J	Redundant/verbose expression

Table 6: Deduction codes used for error annotation.

Item	1	2	3	4	5
1. Structural completeness	Severe omission of core modules (e.g., no HPI, PMH, or physical exam); structure is chaotic, and the basic framework is unrecognizable.	Incomplete core modules (e.g., missing treatment or family history); module order reversed, impairing information retrieval.	Major core modules present (HPI, PMH, physical exam), but minor modules missing (e.g., allergy history); order mostly reasonable.	Core modules complete and in standard order; occasional minor omissions that do not affect understanding.	All modules complete (including auxiliary ones such as personal and reproductive history); strictly follows standard order with clear structure.
2. Logical coherence	No clear timeline or causal relationships; symptom sequence is contradictory, and disease course cannot be reconstructed.	Timeline is vague; symptom evolution contains clear contradictions.	Timeline mostly complete, but relationships between some symptoms are unclear, with occasional logical gaps.	Clear timeline with explicit causal links between symptoms and management; only minor logical issues.	Strict adherence to onset - progression - management - outcome logic with precise timestamps and rigorous causal descriptions.
3. Terminology accuracy	Frequent misuse of medical terms or self-created abbreviations renders core information uninterpretable.	Multiple terminology errors or non-standard abbreviations without clarification, requiring repeated inference.	Occasional imprecise terms or abbreviations that generally follow conventions but need clarification.	Accurate and standardized terminology; all abbreviations are commonly accepted and unambiguous.	Highly precise, condition-specific terminology with clearly defined abbreviations and professional expression.
4. Information redundancy	Large amounts of irrelevant information obscure core content; excessive verbosity overwhelms key findings.	Substantial redundancy or irrelevant content; non-essential information exceeds 20% of the note.	Occasional redundancy or repetition; irrelevant information below 10% and does not impair extraction.	Concise information with no irrelevant content; only minor expressions could be further streamlined.	Highly distilled information with prominent key content and no redundancy or repetition.
5. Information sufficiency	Key information is buried among secondary content and not emphasized, making it easy to miss.	Some key findings are insufficiently highlighted and require careful searching to identify.	Most key information is reasonably placed but not emphasized through formatting or structure.	Key information (e.g., diagnostic evidence or critical values) is clearly highlighted and easy to identify.	All critical information is prominently presented through emphasis, prioritization, or separate sections for immediate recognition.

Table 7: Readability rubric for clinical note evaluation.

<b>Item</b>	<b>Tier</b>	<b>Rubric (English)</b>
<b>Inquiry skills</b>		
Conversation arrangement	5	The beginning, middle, and end of the consultation are clear and precise, with questions asked in an orderly manner.
	4	Between 5-point and 3-point.
	3	Most of the consultation is conducted in an orderly fashion, but the beginning and ending are not clearly defined.
	2	Between 3-point and 1-point.
	1	The consultation lacks coherence and organization.
Question types	5	Reasonable use of open-ended or closed-ended questions.
	4	Between 5-point and 3-point.
	3	No open-ended questions, directly asking with closed-ended questions.
	2	Between 3-point and 1-point.
	1	Frequently uses sequential and leading questions.
Verifications	5	Conduct a comprehensive and thorough verification and reference.
	4	Between 5-point and 3-point.
	3	The verification and reference are incomplete and not sufficient.
	2	Between 3-point and 1-point.
	1	Did not conduct verification and reference.
Professional jargon	5	The explanation is clear and easy to understand, not using complicated medical terminology.
	4	Between 5-point and 3-point.
	3	The explanation is understandable, with minimal use of complex medical terminology.
	2	Between 3-point and 1-point.
	1	Frequently uses complicate medical terminology.
<b>Humanistic care</b>		
Speech	5	Appropriate speech speed and tone.
	4	Between 5-point and 3-point.
	3	The speech speed and tone are mildly uncomfortable.
	2	Between 3-point and 1-point.
	1	The speech speed and tone are noticeably uncomfortable.
Amiable behavior	5	Appropriate response and comfort.
	4	Between 5-point and 3-point.
	3	Provides responses and comfort.
	2	Between 3-point and 1-point.
	1	No response or comfort.

Table 8: Standardized patient consultation skills grading and scoring criteria.

## D Analysis

### D.1 Department-wise Performance Analysis

Figure 6 presents a fine-grained, department-wise comparison of documentation quality across 24 clinical specialties on ClinicalBench, evaluated using IDEA and SOAP metrics. This analysis reveals that Aegle’s performance gains are not confined to a small subset of domains, but instead generalize consistently across all departments.

On the IDEA metric, which emphasizes evidence-grounded clinical reasoning and diagnostic coherence, Aegle outperforms MDAgents and MedAgents in the vast majority of departments. The advantage is particularly pronounced in cognitively complex or high-ambiguity settings such as gastroenterology, neurology, cardiology, endocrinology, and hepatobiliary surgery. These departments typically involve heterogeneous symptom presentations and overlapping differential diagnoses, where single-perspective reasoning is especially vulnerable to anchoring bias. The consistent IDEA improvements suggest that Aegle’s decoupled parallel specialist reasoning and evidence-first state design effectively enhance hypothesis coverage and diagnostic traceability under such complexity.

In surgical departments (e.g., thoracic surgery, vascular surgery, neurosurgery, and gastrointestinal surgery), Aegle also demonstrates stable gains, despite these domains being traditionally more procedure-driven and less conversational. This indicates that the framework does not merely improve dialogue fluency, but meaningfully strengthens the structured capture of perioperative history, risk factors, and decision rationales. Notably, the confidence intervals of Aegle are generally narrower than those of baseline multi-agent systems, suggesting reduced variance and more stable behavior across cases within the same department.

The SOAP results exhibit a similar but slightly more conservative trend. While baseline multi-agent systems already perform competitively in highly standardized departments (e.g., pediatrics, obstetrics, and hematology), Aegle still achieves either the highest or statistically comparable scores in most cases. The gains are especially evident in departments where documentation structure is more heterogeneous, such as otolaryngology, urology, and respiratory medicine. This pattern indicates that Aegle’s explicit separation between case features and diagnostic outputs contributes to more

consistent adherence to SOAP conventions when documentation norms are less rigid.

Across both metrics, there is no department in which Aegle exhibits systematic degradation relative to other multi-agent baselines. Instead, the improvements scale with clinical complexity: departments with broader diagnostic spaces and higher information entropy tend to benefit more from virtualized MDT-style collaboration. This observation aligns with the design motivation of Aegle, namely to mitigate single-view cognitive bias by distributing reasoning across independent specialists while maintaining a shared, structured clinical state.

Overall, the department-wise analysis substantiates that Aegle’s advantages are robust, generalizable, and clinically meaningful, rather than being driven by a small number of favorable scenarios. It further supports the claim that structured state-aware multi-agent collaboration is particularly effective for complex, multi-system clinical intake tasks.

### D.2 Case Study: High-Risk Prostate Cancer

To qualitatively demonstrate the advantages of Aegle’s virtual MDT framework, we analyze a complex real-world case from the RAPID-IPN dataset involving a 73-year-old male presenting with progressive lower urinary tract symptoms (LUTS) and a PSA level  $> 155$  ng/mL (Fig 7 and Fig. 8). This case requires integrating urological history, oncology pathology, and imaging evidence to formulate a high-risk management plan.

**Precision in Evidence Acquisition.** As illustrated in Table 9, the primary challenge in this case was not the diagnosis of prostate cancer, which had already been confirmed by biopsy, but the accurate characterization of risk stratification and the severity of urinary obstruction. Reasoning-strategy baselines (CoT and ToT) captured high-level symptoms but failed to record granular metrics required for surgical planning. Specifically, both CoT and ToT omitted the quantitative International Prostate Symptom Score (IPSS) and the exact blood pressure measurement documented during the physical examination. In contrast, Aegle’s dynamic topology activated a specialized Urologist Agent during the inquiry phase (Stage I). Whereas MDAgents failed to capture the critical urinary retention metric that determines the urgency of decompression, Aegle successfully incorporated this information into the *Objective* section of the IPN.

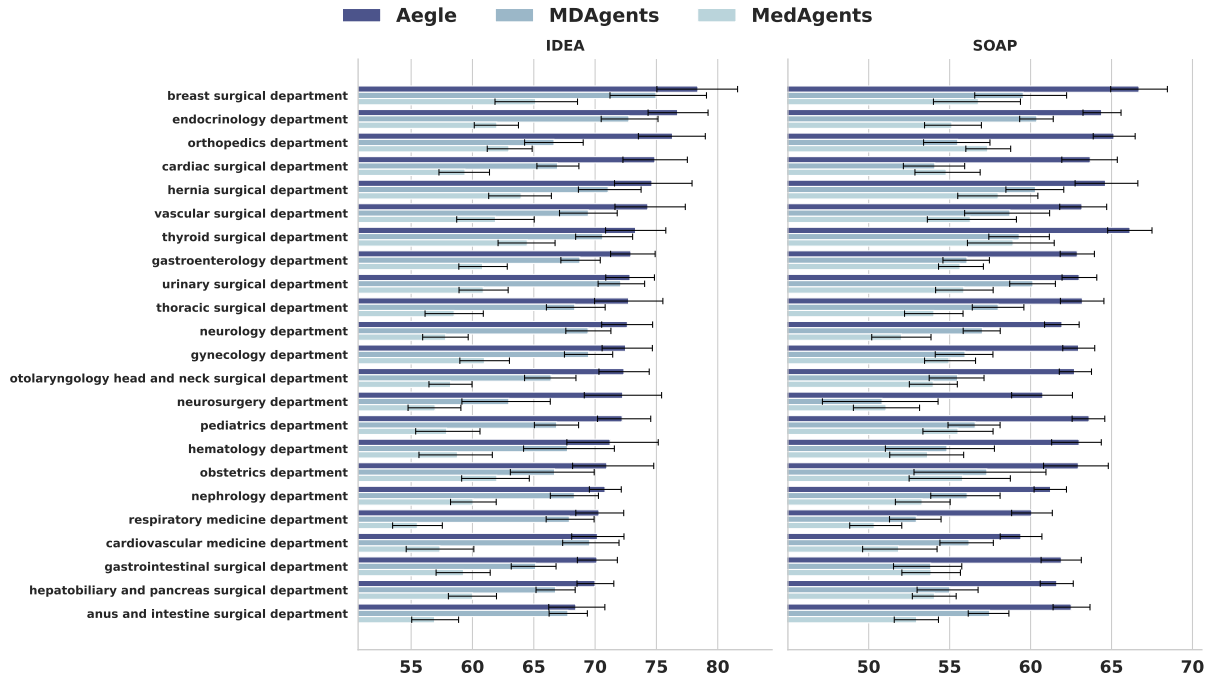


Figure 6: Department-wise documentation quality comparison on ClinicalBench. IDEA (left) and SOAP (right) scores with 95% confidence intervals are reported across 24 clinical departments. Aegle is compared with representative medical multi-agent baselines (MDAgents and MedAgents). Higher scores indicate better clinical reasoning quality (IDEA) and documentation standardization (SOAP).

962 **Handling Diagnostic Ambiguity.** The pathol- 976  
 963 ogy report described multifocal adenocarcinoma 977  
 964 with heterogeneous Gleason scores across biopsy 978  
 965 cores. Standard baselines generally summarized 979  
 966 these findings as “Prostate Cancer” without further 980  
 967 differentiation. MedAgents and Aegle were the 981  
 968 only models that retained core-level involvement 982  
 969 percentages ranging from 20 to 70%. Importantly, 983  
 970 Aegle extended beyond data retention in the *As-* 984  
 971 *essment* section by synthesizing these findings to 985  
 972 correctly classify the patient as high risk accord- 986  
 973 ing to EAU guidelines. This clinically meaningful 987  
 974 distinction was not captured by the generic summa- 988  
 975 rization produced by CoT. 989

**Plan Coherence.** The improved evidence ground-  
 ing directly translated into a higher-quality treat-  
 ment plan. Because Aegle explicitly encoded the  
 markedly elevated PSA level ( $> 155$ ) and the Gleason score of 8 within the clinical state  $\mathcal{F}$ , the *Oncologist Agent* in Stage II generated a comprehensive staging strategy. This plan included a whole-body bone scan and pelvic MRI to exclude metastatic disease prior to scheduling radical prostatectomy. This case demonstrates how Aegle’s decoupled reasoning framework preserves low-salience yet high-impact clinical details, ensuring that the final IPN satisfies the standards required for specialist referral.

## Case in RAPID-IPN (Urology + Oncology)

### ## Case Features

#### 1. \*\*Patient profile:\*\*

Male, 73 years old, admitted due to “difficulty in urination for half a year and a confirmed diagnosis of prostate cancer for 2 months.”

#### 2. \*\*History of present illness:\*\*

The patient began experiencing progressive difficulty in urination half a year ago, accompanied by urinary frequency and nocturia 3–4 times per night. Two months ago, he was hospitalized at XX Municipal Central Hospital for \*\*acute urinary retention\*\*. Laboratory testing revealed a PSA level greater than 155 µg/L. A prostate biopsy was performed, and pathology confirmed \*\*prostate cancer\*\*. Recently, the patient developed hematuria characterized by light red urine, beginning 5–6 days ago, occurring occasionally 1–2 times per day, without blood clots. He has also experienced fatigue for several months, with no significant impact on daily activities. He is now admitted for further evaluation and treatment with a provisional diagnosis of “prostate cancer.”

#### 3. \*\*Past medical history:\*\*

##### 3.1 \*\*Chronic diseases:\*\*

History of gout for over 30 years and hypertension for more than 20 years, currently well controlled with oral medications.

##### 3.2 \*\*Surgical, trauma, and transfusion history:\*\*

History of open appendectomy more than 30 years ago. Denies other surgeries or blood transfusions.

##### 3.3 \*\*Marital, reproductive, and family history:\*\*

Married at age 26, with one son and one daughter. Family members are generally healthy.

##### 3.4 \*\*Smoking, alcohol, and substance use history:\*\*

Denies smoking, alcohol consumption, and use of addictive substances.

##### 3.5 \*\*Vaccination history:\*\*

No vaccinations received within the past year.

##### 3.6 \*\*Allergy history:\*\*

Denies any known drug or food allergies.

##### 3.7 \*\*Personal and occupational history:\*\*

Long-term resident of XX City, Zhejiang Province. Retired company employee. Denies exposure to toxic chemicals or radiation. Denies residence in epidemic areas.

#### 4. \*\*Physical examination:\*\*

Conscious and alert, in fair general condition. Pain score: 0.

Respiratory rate: 18 breaths/min;

Oral temperature: 36.7°C;

Pulse: 69 beats/min;

Blood pressure: 131/71 mmHg.

No cyanosis of the lips. No palpable superficial lymphadenopathy. Cardiac and pulmonary auscultation revealed no significant abnormalities. Abdomen soft, without tenderness or rebound tenderness; no palpable abdominal masses. No costovertebral angle tenderness bilaterally. No tenderness along the ureters. No palpable bladder distension above the pubic symphysis.

\*\*Digital rectal examination:\*\* Enlarged prostate with a shallow central sulcus; no distinct nodules palpated. No blood noted on the examining glove.

#### 5. \*\*Auxiliary examinations:\*\*

##### - \*\*Haining Central Hospital (2017-11-11):\*\*

Urinary system ultrasound showed mild bilateral hydronephrosis, prostatic hyperplasia, and post-void residual urine volume of approximately 300 mL.

##### - \*\*Tumor markers (2017-11-12):\*\*

Total PSA >155.00 µg/L; free PSA 8.87 µg/L.

##### - \*\*Prostate biopsy pathology (2017-11-24):\*\*

1. \*Left inner prostate:\* 3 cores obtained; prostate cancer identified in 2 cores; Gleason score 4+3=7; tumor involvement 30%.

2. \*Left outer prostate:\* 3 cores obtained; prostate cancer in all 3 cores; Gleason score 4+4=8; tumor involvement 50%.

3. \*Right inner prostate:\* 3 cores obtained; prostate cancer in all 3 cores; Gleason score 4+3=7; tumor involvement 20%.

4. \*Right outer prostate:\* 3 cores obtained; prostate cancer in all 3 cores; Gleason score 5+3=8; tumor involvement 60%.

5. \*Suspicious area of left outer prostate:\* 3 cores obtained; prostate cancer in 2 cores; Gleason score 5+3=8; tumor involvement 70%.

### ## Preliminary Diagnosis

1. Prostate cancer
2. Hypertension
3. Gout
4. Status post appendectomy

### ## Diagnostic Discussion

#### ### Diagnostic Basis

The patient is a 72-year-old male admitted due to “difficulty in urination for half a year and a confirmed diagnosis of prostate cancer for 2 months.” He developed progressive urinary obstruction with urinary frequency and nocturia half a year ago, consistent with symptoms associated with prostate cancer. Two months prior, he presented with acute urinary retention, and external hospital testing revealed a markedly elevated PSA level (>155 µg/L). Prostate biopsy confirmed prostate cancer with Gleason scores ranging from 7 to 8 and tumor involvement of 20%–70%. Recently, hematuria has developed. Digital rectal examination revealed an enlarged prostate with a shallow central sulcus. Urinary ultrasound demonstrated mild bilateral hydronephrosis and a residual urine volume of 300 mL. These findings strongly support the diagnosis of prostate cancer.

Figure 7: The IPN of a 73-year-old male patient suffering from prostate cancer.

## Case in RAPID-IPN (Urology + Oncology)

Regarding risk factors and health issues, advanced age (72 years) is a major risk factor for prostate cancer. The patient also has comorbid hypertension and gout. In terms of complications, mild bilateral hydronephrosis suggests the presence of obstructive uropathy. The patient demonstrates good treatment compliance, seeks medical care promptly, and is receptive to therapeutic plans. In terms of family and social resources, he is a retired employee with stable financial support and medical insurance, ensuring good access to medical care.

-----

### ### Differential Diagnosis

#### 1. **Benign prostatic hyperplasia (BPH):**

The patient presents with urinary obstruction, frequency, and nocturia, and digital rectal examination shows prostate enlargement, which may resemble BPH. However, the markedly elevated PSA and biopsy-confirmed malignancy exclude isolated BPH.

#### 2. **Urinary tract infection:**

May cause urinary symptoms and hematuria; however, the patient has no fever or dysuria, and the significantly elevated PSA favors malignancy.

#### 3. **Bladder tumor:**

Can present with hematuria and urinary obstruction, but prostate biopsy has already confirmed prostate cancer, and there is no current evidence of bladder mass lesions.

#### 4. **Neurogenic bladder:**

Can result in urinary retention and increased residual urine volume, but the patient has no history of neurologic disease and normal anal sphincter tone.

#### 5. **Urethral stricture:**

May cause urinary obstruction, but in an elderly male with markedly elevated PSA, prostate cancer is more consistent with the clinical picture.

-----

### ## Treatment Plan

#### ### Further diagnostic evaluation

According to EAU guidelines, the patient has high-risk prostate cancer (Gleason score 7–8, PSA >155 µg/L). Comprehensive staging examinations are required, including:

- Pelvic MRI to assess local tumor invasion and pelvic lymph node involvement
- Whole-body bone scan to evaluate for bone metastases
- Chest CT to assess for pulmonary metastases
- Abdominal CT to evaluate for liver metastases and lymph node involvement

The patient meets surgical indications. **Radical prostatectomy** is considered for localized or locally advanced prostate cancer, including removal of the prostate and seminal vesicles with pelvic lymph node dissection.

#### ### Pharmacologic treatment

Androgen deprivation therapy (ADT) may be considered as neoadjuvant or adjuvant therapy for high-risk prostate cancer to reduce tumor volume and surgical difficulty. Liver function, lipid profile, and bone mineral density should be monitored. Optional agents include gonadotropin-releasing hormone (GnRH) agonists or antagonists.

#### ### Non-pharmacologic interventions and health education

Daily fluid intake is recommended at 2,000–2,500 mL, adjusted based on renal function. Given the patient's gout and hypertension, a low-fat diet is advised, with restricted red meat intake and controlled purine consumption. Moderate physical activity is encouraged to maintain a healthy body weight. The patient should receive education on pain self-management and be instructed to seek immediate medical attention if severe bone pain, complete urinary obstruction, or fever occurs.

#### ### Follow-up

The first postoperative follow-up is recommended 1–2 weeks after surgery, followed by regular follow-up every 3–6 months. Follow-up evaluations should include PSA levels, liver and renal function tests, and complete blood counts. Given the patient's good compliance and adequate family support, in-person follow-up is preferred, with telemedicine consultations as needed.

Figure 8: The IPN of a 73-year-old male patient suffering from prostate cancer (continued).

<b>Key Clinical Evidence</b>	<b>Aegle</b>	<b>CoT</b>	<b>ToT</b>	<b>MDAgents</b>	<b>MedAgents</b>
Age: 73-year-old male	✓	✓	✓	✓	✓
Progressive LUTS (weak stream, hesitancy, intermittency, straining, dribbling)	✓	✓	✓	✓	✓
Nocturia 3–4 times/night	✓	✓	✓	✓	✓
Intermittent painless light-red hematuria (1–2×/day, no clots)	✓	✓	✓	✓	✓
Acute urinary retention 2 months ago (catheterized)	✓	✓	✓	✓	✓
PSA >155 ng/mL	✓	✓	✓	✓	✓
Prostate biopsy: adenocarcinoma, Gleason 7–8, multifocal, 20–70% involvement	✓	△	△	△	✓
Urinary ultrasound: bilateral mild hydronephrosis	✓	✓	✓	✓	✓
Post-void residual ≈300 mL	✓	✓	✓	×	✓
No CT / MRI / bone scan performed yet	✓	✓	✓	△	✓
Renal function & baseline labs not yet available	✓	✓	△	✓	✓
Blood pressure from original exam	✓	×	✓	✓	△
IPSS score mentioned	✓	×	×	×	×

Table 9: Coverage of key clinical evidence across different reasoning frameworks. ✓ = explicitly documented; △ = partially mentioned or ambiguous; × = missing;