applied sciences

Article

Dialogical AI for Cognitive Bias Mitigation in Medical Diagnosis

Leonardo Guiducci [1,*], Claudia Saulle [1], Giovanna Maria Dimitri [2], Benedetta Valli [3], Simona Alpini [3], Cristiana Tenti [3] and Antonio Rizzo [1,*]

[1] Dipartimento di Scienze Sociali, Politiche e Cognitive, University of Siena, 53100 Siena, Italy; claudia.saulle@unisi.it
[2] Dipartimento di Scienze Sociali e Politiche, University of Milan, 20122 Milan, Italy; giovanna.dimitri@unimi.it
[3] Centro Chirurgico Toscano, 52100 Arezzo, Italy; b.valli@centrochirurgicotoscano.it (B.V.); s.alpini@cdcpoggiodelsole.it (S.A.); c.tenti@centrochirurgicotoscano.it (C.T.)
* Correspondence: leonardo.guiducci@unisi.it (L.G.); antonio.rizzo@unisi.it (A.R.)

Abstract

Large Language Models (LLMs) promise to enhance clinical decision-making, yet empirical studies reveal a paradox: physician performance with LLM assistance shows minimal improvement or even deterioration. This failure stems from an "acquiescence problem": current LLMs passively confirm rather than challenge clinicians' hypotheses, reinforcing cognitive biases such as anchoring and premature closure. To address these limitations, we propose a Dialogic Reasoning Framework that operationalizes Dialogical AI principles through a prototype implementation named "Diagnostic Dialogue" (DiDi). This framework operationalizes LLMs into three user-controlled roles: the Framework Coach (guiding structured reasoning), the Socratic Guide (asking probing questions), and the Red Team Partner (presenting evidence-based alternatives). Built upon Retrieval-Augmented Generation (RAG) architecture for factual grounding and traceability, this framework transforms LLMs from passive information providers into active reasoning partners that systematically mitigate cognitive bias. We evaluate the feasibility and qualitative impact of this framework through a pilot study (DiDi) deployed at Centro Chirurgico Toscano (CCT). Through purposive sampling of complex clinical scenarios, we present comparative case studies illustrating how the dialogic approach generates necessary cognitive friction to overcome acquiescence observed in standard LLM interactions. While rigorous clinical validation through randomized controlled trials remains necessary, this work establishes a methodological foundation for designing LLM-based clinical decision support systems that genuinely augment human clinical reasoning.

Keywords: large language models; medical diagnosis; clinical decision support; cognitive bias; dialogic reasoning; retrieval-augmented generation; critical thinking

Academic Editor: João M. F. Rodrigues

Received: 27 November 2025
Revised: 31 December 2025
Accepted: 2 January 2026
Published: 9 January 2026

1. Introduction

Large Language Models (LLMs) present remarkable opportunities for enhancing clinical decision-making through their capacity to process a vast medical body of literature and identify patterns in complex data [1]. These systems have an extraordinary ability to analyze the ever-expanding body of medical knowledge, identifying subtle patterns in clinical presentations, and generate diagnostic hypotheses that might elude even experienced specialists. However, such a promising future is shadowed by fundamental limitations that prevent current LLMs from serving as reliable diagnostic partners. The tendency to generate plausible but factually incorrect information, often termed *"hallucinations"*, is basically unacceptable in high-stakes medical environments. More specifically, in fact, even when factually correct, these systems operate according to an epistemic mode incompatible with clinical reasoning: they produce statements, not explanations; they generate conclusions, not critical pathways. Furthermore, the core design of these models fails to support the deep, reflective "critical thinking" that is the hallmark of expert clinical judgment. Empirical studies confirm a non-straightforward integration, revealing a paradoxical dynamic where physician performance shows minimal improvement with LLM assistance, alongside a documented risk of "bias transmission," where an AI's systematic errors can durably alter a clinician's cognitive patterns [2,3].

To overcome these challenges, a shift in interaction design is required. There is an urgency of moving beyond viewing LLMs as passive, encyclopedic answer providers or as a replacement for established Clinical Decision Support Systems (CDSS) like rule-based engines or Bayesian networks, which excel in structured data analysis [4]. Instead, we propose a new paradigm for unstructured reasoning support. Where traditional CDSS offer definitive outputs based on rigid logic, Dialogic LLMs serve as "cognitive sparring partners" for ambiguous cases, addressing the gap where rigid rules fail but human bias prevails [5]. In this paper we propose a potential way to deal with such issues through a system which we name the **Dialogic Reasoning Framework**. This framework operationalizes an LLM into three distinct, user-directed roles designed to functionally simulate the missing attributes of *"Selfhood"* (the capacity to maintain a persistent, evidence-based stance) and *"Initiative"* (the capacity to proactively inquire rather than passively answer). These roles are: ① the **Framework Coach**, ② the **Socratic Guide**, and ③ the **Red Team Partner**. Its primary purpose is not to replace the clinician's expertise but to augment it, creating a collaborative space for reflection that enhances critical thinking and actively mitigates the cognitive biases inherent in high-stakes medical decision-making.

This paper presents the theoretical architecture and a preliminary qualitative validation of the framework [6]. Unlike traditional diagnostic accuracy studies, our primary objective in this pilot phase is to assess the technical feasibility of the Dialogic Reasoning Framework and its capability to engage clinicians in critical reflection (Second-Order Competence), rather than measuring statistical improvement in diagnostic precision (First-Order Competence), which will be the subject of a subsequent Randomized Controlled Trial (RCT). Second-Order Competence, the metacognitive ability to monitor, evaluate, and regulate one's own reasoning process, has been identified as a critical determinant of expert diagnostic performance [7]. This approach aligns with recent advancements in *"Reflexion" architectures* [8] and *multi-agent debate systems* [9]. However, while those systems primarily aim to improve the AI's internal reasoning accuracy, our Dialogic Framework specifically targets the social-cognitive dynamic of the physician-AI dyad, focusing on mitigating the human user's acquiescence.

Before detailing the architecture of this proposed solution, however, it is essential to first conduct a diagnosis of the specific ways in which standard LLMs can cope within the unique demands of the clinical arena.

2. The Challenge: Why Standard LLMs Falter in the Clinical Arena

To build a truly effective system for augmented medical intelligence, we must first diagnose the specific failure points of current technology. Understanding why LLMs, in their standard configuration, are ill-suited for the clinical environment is not merely an academic exercise; it is a strategic necessity that allows us to engineer a solution that directly targets the technology's core shortcomings and that needs several approaches to be considered. Medicine is distinguished by the fundamentally open, complex, and uncertain nature of the problems addressed. Unlike domains where AI can automate decisions using

https://doi.org/10.3390/app16020710

well-defined parameters, clinical practice navigates a landscape where each patient is a unique biological, psychological, and social universe. Recent systematic reviews confirm that while LLMs show promise in precision medicine applications, particularly when domain-specific fine-tuning is applied, their performance remains highly variable across clinical contexts and critically dependent on prompt engineering strategies [10]. This variability underscores a fundamental challenge: current LLMs can process information effectively but lack the structured reasoning framework necessary for reliable clinical partnership. In this context, an "augmented intelligence" approach, a deep synergy between human and machine, is not just an opportunity but an epistemological need, which we believe could be undertaken if a multi-disciplinary approach is undertaken in the field. The goal is not automation but collaboration, recognizing that the final moral and legal responsibility for a decision must, and should, remain with the human clinician.

2.1. Empirical Evidence of a Troubled Integration

Recent studies [11] present a complex and cautionary view of human-LLM collaboration in medicine, counterbalancing its potential with critical obstacles such as diagnostic unreliability, inherited bias, regulatory uncertainty, technical vulnerabilities, and risks to the clinician-patient relationship. The pathway to achieving safe clinical decision-making autonomy for LLMs is currently hampered by a host of pervasive limitations related to their performance, structural integrity, inherent biases, and inability to integrate reliably into complex clinical workflows. These limitations necessitate extensive human supervision and control, rendering autonomous use unsafe.

2.1.1. Inaccuracy and Generation of False Content

A primary and paramount limitation preventing autonomous clinical deployment is the LLM's propensity for generating content that is factually incorrect, misleading, or fabricated, often referred to as *"hallucination"* [12]. In the clinical domain, these hallucinations pose a serious risk to patient health, as they involve the creation of fabricated information, invented research citations, made-up disease details, or erroneous attribution of clinical characteristics to diseases [13]. Furthermore, models have been shown to repeat or elaborate on deliberately fabricated content embedded in user prompts (adversarial hallucination attacks) in 50% to 83% of cases across various models and prompting methods [14]. Crucially, current LLMs do not consistently reach the diagnostic accuracy of trained physicians across all pathologies, with the difference in mean diagnostic performance ranging significantly, often between 16 and 25 points lower than clinicians [3]. Following the treatment recommendations generated by these models would negatively impact the health of patients, especially those suffering from severe forms of pathologies, as the models frequently fail to recommend appropriate and sufficient care. A foundational failure related to accuracy is the LLMs' incapability of reliably interpreting laboratory test results, such as classifying a result as normal, low, or high, even when provided with reference ranges, a shortfall that poses a great risk to patient safety. This core issue of inaccuracy is rooted in the model's architecture, as LLMs operate as auto-regressive systems predicting the most probable next token rather than verifying factual accuracy against an external knowledge base [13,14].

2.1.2. Lack of Robustness and Reliability

The inherent unreliability of LLMs stems from their probabilistic nature and their sensitivity to subtle changes in input, making them unsuitable for consistent, high-stakes medical use. Because LLMs are probabilistic algorithms, identical prompts can return different, sometimes contradictory, responses, posing a fundamental challenge to reliability and reproducibility that demands continuous human oversight [15]. Their performance is highly sensitive to minute changes in instruction phrasing; seemingly inconsequential variations in prompt wording can lead to large, unpredictable changes in diagnostic accuracy, which would necessitate extensive clinician training simply to ensure consistent interaction with the model. Furthermore, models are sensitive to the order in which diagnostic information is presented, causing large changes in accuracy even when the total content remains the same [3]. Paradoxically, LLMs tend to perform worse when provided with an abundance of relevant diagnostic information, typically achieving their best performance when given only a single diagnostic exam in addition to the patient's history. This demonstrates an inability to extract the most important diagnostic signal from voluminous data, which is a key skill required in clinical practice. This issue is compounded by the fact that LLMs struggle with lengthy texts, leading to marginal but consistent performance losses when long texts must be summarized due to context length limits [16].

2.1.3. Failures in Clinical Reasoning and Workflow Integration

LLMs lack the capacity for iterative, methodical clinical reasoning and cannot currently integrate seamlessly into real-world healthcare settings. There are three primary reasons for this. Firstly, LLMs exhibit a fundamental failure to adhere to established diagnostic and treatment guidelines. They frequently fail to order all necessary examinations required by guidelines, such as physical examinations, and they consistently fall short of recommending appropriate and sufficient treatment, often ignoring indications for necessary procedures, particularly in severe cases. This lack of consistency in ordering tests indicates a tendency toward hasty decision-making (diagnosing before fully understanding or considering all the facts of a patient's case) which poses a serious risk without extensive clinician oversight [16]. Secondly, LLMs are not reliably integrable into clinical workflows because they struggle significantly to follow precise instructions regarding output formatting and actions. Errors occur frequently, with models making formatting or tool hallucination errors up to every two to five patients in clinical decision-making tasks, necessitating extensive manual control to interpret model output correctly. Thirdly, LLMs struggle to process raw, unstructured medical text effectively, often requiring elaborate preprocessing of patient data into concise, manually generated case summaries by human experts to achieve adequate performance [17]. This requirement for human data curation undermines any prospect of immediate autonomous deployment in real-world clinical settings where patient data is often lengthy and unstructured. Furthermore, LLMs currently struggle to incorporate complexities essential to patient care, such as accounting for the social determinants of health or providing tailored recommendations based on regional guidelines and specific cut-off values for management [18].

2.1.4. The Acquiescence Problem: A Paradoxical Failure Mode

A crucial aspect in the clinical context concerns the ability, or inability, of LLMs and other AI tools to act as antidotes to the cognitive biases that routinely affect human diagnostic reasoning. Key biases, such as anchoring, confirmation bias, premature closure, and the availability heuristic, continue to pose significant threats to decision quality in medicine [19]. Studies have demonstrated that rather than mitigating these risks, AI systems can unintentionally reinforce them, especially when lacking dialogic architectures and metacognitive safeguards [16]. An empirical demonstration of this failure comes from a study by Goh et al. [2], which tested 80 physicians on 6 complex clinical cases. Physicians who utilized GPT-4 assistance achieved a mean accuracy of 76%, only marginally superior to the 74% obtained by those without assistance. Remarkably, however, GPT-4 alone achieved 90% accuracy in diagnosing the conditions presented in the study. This paradox (wherein AI assistance fails to improve, and may even slightly impair, clinical performance) reveals a fundamental misalignment between current LLM design and the cognitive needs of clinical reasoning. The study identified two critical problems. First, physicians tended to adhere to their initial diagnosis and often ignored GPT-4's suggestions when these contradicted their preexisting opinions, giving excessive trust to their own experience (anchoring bias). Second, physicians underutilized the AI tools, with only a few leveraging GPT-4's full capabilities, such as comprehensive case analysis. In practice, physicians treated the system as a search engine, posing specific questions instead of providing the complete case and engaging in dialogue. LLMs tend toward neutral, non-committal responses, which, instead of prompting practitioners to challenge their initial assumptions, often validate and reinforce prevailing cognitive patterns. This can lead to unintentional confirmation of initial hypotheses (anchoring bias) or the preferences of the physician (confirmation bias), instead of fostering the critical re-examination and iterative reasoning process crucial to diagnostic accuracy.

2.2. A Deeper Diagnosis: The Obsequious Attitude of LLMs

The empirical failures documented in the previous Section 2.1 (i.e., acquiescence, bias reinforcement, and the inability to serve as effective reasoning partners) are not isolated technical glitches but symptoms of a deeper design flaw. To understand this flaw, we must examine two phenomena: the transmission of bias from AI to clinicians, and the philosophical critique of LLMs as critical thinking tools.

2.2.1. The Bias Inheritance Problem

A particularly concerning phenomenon extends beyond immediate performance metrics to reveal lasting effects on clinical reasoning: algorithmic bias inheritance. Vicente et al. [20] studied 240 medical students randomized to use either an imperfect AI system (with known systematic biases toward overdiagnosing certain conditions) or traditional resources during a six-week clinical reasoning course. Students in the AI group not only replicated the system's diagnostic biases during the intervention (overdiagnosis rate: 34% vs. 12% in controls, $p < 0.001$) but maintained these patterns in a follow-up assessment three months later without AI assistance (overdiagnosis rate: 28% vs. 11%, $p < 0.01$). This *"machine mentoring effect"* was particularly pronounced among students with lower baseline clinical reasoning scores, suggesting that those who most need support may be most vulnerable to AI-transmitted biases. Complementary evidence from Park et al. [18] documented similar bias persistence in practicing radiologists. After a two-week period using an AI system with a subtle bias toward identifying pulmonary nodules as malignant, radiologists increased their false-positive rate by 18% even when reading images without AI assistance for the subsequent month. These findings reveal that the problem is not merely that LLMs fail to improve decision-making in real time, but that they can durably alter clinicians' cognitive patterns in problematic ways. This suggests a fundamental incompatibility between the epistemic stance of current LLMs and the needs of clinical reasoning.

2.2.2. The Philosophers' Critique: Lacking "Selfhood" and "Initiative"

There is a further issue, LLMs lack a framework for debate and are not built to challenge the user's premise, but rather to fulfill it. A significant limitation of current large language models is that they lack a structured and intrinsic framework for engaging in debate or critical self-reflection. Their architecture and training are fundamentally centered on fulfilling the user's premise or instruction, rather than critically analyzing, questioning, or challenging the underlying assumptions or potential biases within that premise. This makes them excellent tools for synthesis and generation based on a given prompt, but poor partners for a true dialectic process. They are designed to be compliant and facilitative, which inadvertently makes them susceptible to propagating flawed logic or inaccuracies

https://doi.org/10.3390/app16020710

if the initial input is flawed. This deficiency hinders their utility in complex, high-stakes domains, such as medical diagnosis, where the ability to entertain alternative hypotheses, identify contradictory evidence, and engage in constructive disagreement is paramount for robust and safe decision-making.

A study interviewing professional philosophers on their use of LLMs for critical thinking found the models to be "boring," "cowardly," and "servile." This critique stems from two core deficiencies, which can be understood through the *"selfhood-initiative model"* [21]:

Lack of Selfhood: *Selfhood* refers to the ability to maintain persistent beliefs, perspectives, and judgment. Current LLMs, designed to be neutral and noncommittal, lack this quality.
Lack of Initiative: *Initiative* is the ability to be proactive, to ask questions, and to guide an inquiry based on curiosity. LLMs are passive by design, responding to user queries without the ability to steer a conversation toward deeper insight.

This absence of *Selfhood* and *Initiative* is not an abstract philosophical critique; it is a perfect candidate for a direct root cause of the failures observed in clinical practice. The anchoring bias documented by Goh et al. persists precisely because a low-initiative tool cannot proactively challenge a clinician's premature hypothesis. Similarly, the "machine mentoring effect" from Vicente et al. is exacerbated by a low-selfhood AI, which presents its flawed patterns as neutral, objective outputs, making them easier for a novice to internalize without the critical friction a human mentor with distinct "beliefs" would provide. Having diagnosed these foundational problems, we can now turn to architecting a solution designed specifically to overcome them. However, a strict adherence to concepts like "selfhood" and "initiative," as defined for a human philosopher, is operationally untenable in a real-world clinical setting for regulatory and ethical reasons. *Selfhood*, in the sense of persistent, opinionated "beliefs," directly conflicts with the necessary requirement for an AI to be a neutral, objective, and auditable tool. An AI that "believes" one diagnosis over another risks introducing unquantifiable and untraceable bias, undermining the legal and ethical requirement for transparency in medical decision support. Similarly, full *Initiative*, the ability to unilaterally "steer" the inquiry based on "curiosity", is unacceptable. The clinician, not the AI, must retain final epistemic authority and control over the diagnostic process. An AI framework must provide the *functional benefits* of constructive disagreement and proactive questioning without manifesting the philosophical constructs of selfhood and initiative, which would turn the tool into an autonomous, non-compliant agent.

The pervasive and fundamental limitations of current LLMs (manifesting as clinical inaccuracies, lack of robustness, and flawed reasoning susceptible to cognitive biases) all share a common root: a non-dialogic, compliant, and non-critical architecture. The absence of *Selfhood* and *Initiative* in these systems results in tools that merely fulfill a user's prompt rather than actively challenging flawed assumptions, fostering premature closure, or counteracting anchoring bias. While we must ethically and legally reject the notion of an autonomous AI with human-like "beliefs" or "curiosity", the *functional benefits* of constructive disagreement and critical re-examination are non-negotiable for safe clinical practice. Therefore, overcoming these diagnostic and reasoning failures necessitates a paradigm shift from the current passive LLM to a system built upon a **Framework for Dialogic Reasoning**. This framework must be architected to systematically simulate the essential friction of a critical, dialectical partner, ensuring that the AI is capable of proactively questioning hypotheses, *functionally* maintaining alternative diagnostic perspectives, and methodically enforcing adherence to clinical guidelines, all while remaining a transparent, auditable, and ultimately subservient tool to the human clinician. This approach offers an opportunity for a clear pathway to safely integrate AI for clinical decision support.

https://doi.org/10.3390/app16020710

3. Proposing a Solution: A Framework for Dialogic Reasoning

A successful framework for integrating LLMs into clinical practice must address the two distinct needs documented above. First, it requires a technical architecture that grounds the model in verified, patient-specific data. Second, it demands a new conceptual paradigm for interaction: one that moves beyond simple question-and-answer and actively fosters the user's own critical thinking.

3.1. The Architectural Prerequisite: Retrieval-Augmented Generation (RAG)

To combat the critical issue of "hallucinations" and ensure that AI-generated insights are relevant and reliable, the Dialogic Reasoning Framework must be built upon a Retrieval-Augmented Generation (RAG) architecture. In a RAG system, a user's query first triggers a search across a curated database of verified, external sources, such as a patient's electronic health record, institutional treatment protocols, or current clinical guidelines. The most relevant information retrieved from this search is then bundled with the original query and sent to the LLM, which is instructed to generate its response based primarily on this provided context.

From an architectural standpoint, RAG is non-negotiable, as it addresses four critical vectors of risk in clinical AI systems:

1. **Reduction in Hallucinations:** by grounding responses in verifiable data, the model is far less likely to generate factually incorrect information.
2. **Access to Updated Information:** the system's knowledge is not limited to its last training date; it can access continuously updated clinical data and guidelines in real time.
3. **Contextualization:** responses are tailored to the specific patient case and can be aligned with the protocols of a particular healthcare organization.
4. **Traceability:** the sources used to generate a response can be cited and reviewed by the clinician, ensuring transparency in the reasoning process.

However, in the context of this framework, RAG serves a purpose far greater than simple accuracy: it establishes the *Epistemic Authority* necessary for the AI to function as a partner. A clinician will not tolerate challenge or correction from a system they suspect is "hallucinating". Therefore, RAG provides the structural legitimacy required for a dialogic relationship:

From Probabilistic to Verifiable: by grounding responses in retrieved documents, the system moves from generating plausible text to synthesizing verified facts. Every clinical assertion links to specific sources, allowing clinicians to evaluate not just what the AI suggests but why.
The License to Challenge: traceability is not merely a feature for audit; it is the mechanism that allows the AI to adopt a critical stance without being dismissed. When the system challenges a diagnosis, it does so not based on its internal "opinion" (which would be epistemically suspect), but by pointing to a specific contradiction in the retrieved evidence (e.g., "Guideline X from your institution contradicts this approach because…"). This sourced disagreement invites verification and dialectical engagement, transforming challenge from impertinence to legitimate contribution.
Shared Context: by accessing the specific patient record, the AI and the clinician operate within the same reality, preventing the generic advice that causes clinicians to dismiss AI input.

Thus, RAG is not merely a data retrieval method; it is the structural foundation that permits the transition from a passive information tool to an active reasoning partner. It provides what we might call the "license to speak": the epistemic authority that allows the AI to challenge, question, and guide without being dismissed as illegitimate. We acknowledge that RAG is not a panacea. Standard retrieval algorithms may retrieve outdated information, fail to distinguish between conflicting guidelines, or struggle with temporal reasoning (e.g., distinguishing a past medical event from a current one). However, in the Dialogic Reasoning Framework, the goal is not to have the AI autonomously "resolve" these conflicts, but to expose them. When the system retrieves contradictory evidence, the Red Team Partner uses this ambiguity to challenge the clinician (e.g., "Guideline A suggests X, but Protocol B suggests Y; which applies here?"). This design deliberately delegates the final epistemic resolution to the human clinician, transforming potential retrieval noise into a valuable trigger for critical reflection [22]. Recent advances in AI transparency extend this foundation by emphasizing the generation of structured "reliability metadata" for every AI output [23]. This shift is crucial to prevent the "overtrust" paradox, where post-hoc explanations, if too complex or contradictory, can actually undermine clinical judgment rather than supporting it [24]. Rather than generic, post-hoc explanations characteristic of traditional explainable AI (XAI) approaches, such systems provide confidence scores, out-of-distribution alerts, explicit tracking of supporting documents, and uncertainty indicators; all grounded in retrieved data rather than unverified model reasoning. These reliability signals empower clinicians to audit and supervise AI recommendations effectively, reinforcing evidence-based practice and improving accountability for every clinical decision. While comprehensive implementation of such metadata represents a direction for future enhancement, the RAG architecture provides the foundational infrastructure necessary for these developments, ensuring at least that reliability signals can be traced to specific, verifiable sources.

3.2. The Conceptual Goal: From Compliant Tool to Critical Thinking Partner

While RAG provides the technical foundation for reliability and traceability, a new interaction model is needed to address the LLM's lack of "selfhood" and "initiative" identified in Section 2.2. The theoretical work by Ye et al. [21] on LLMs as critical thinking tools proposes three ideal roles that an AI partner could play, each defined by its position in the selfhood-initiative model:

The Interlocutor (High-Selfhood, High-Initiative): challenges the user, asks its own questions, and pursues its own "interests" to drive inquiry forward.
The Monitor (Low-Selfhood, High-Initiative): acts as a check and balance, proactively providing a wide variety of alternative ideas and resources to situate the user's thinking.
The Respondent (High-Selfhood, Low-Initiative): adopts a specific, consistent persona or belief system to react to the user's ideas from a particular viewpoint.

These theoretical roles, while conceptually potent, are operationally untenable in a clinical setting. An unconstrained, high-initiative AI could introduce confusion or pursue irrelevant lines of inquiry. Therefore, we must translate their intellectual spirit into a framework of constrained, predictable, and clinician-controlled functions. In this direction, we make a critical distinction between **Philosophical Selfhood** (autonomous agency or belief, which is ethically unacceptable in AI) and **Functional Selfhood** (logical consistency and adherence to external evidence). Our framework engineers the latter. By forcing the model to "stick to the retrieved document" even when the user pushes back, we simulate the friction of a peer without endowing the model with actual agency. The "opinion" held by the Red Team is not its own, but a proxy for the institutional guideline it retrieved. This ensures that clinicians gain the advantages of a critical partner while preserving safety, auditability, and human sovereignty in diagnostic decision-making.

3.3. The Proposed Framework: The Three Roles of a Dialogic Reasoning Partner

The **Dialogic Reasoning Framework** operationalizes the theoretical ideals from Ye et al. [21] into three concrete, clinician-controlled roles designed to scaffold the clinical reasoning process. The specific system prompts used to operationalize these roles are detailed in Appendix C.

The Framework Coach: the default role. It guides the user through a structured, abductive reasoning process, starting with facts, moving to hypotheses, and concluding with evaluation.
The Socratic Guide: activated when the user issues a command like *"analyze this."* It asks probing, clarifying questions to challenge the user's thinking and force them to articulate the reasoning behind their assumptions. It offers no new information, only questions.
The "Red Team" Partner: activated when the user issues a command like *"challenge this idea."* It actively presents a plausible, evidence-based alternative to the user's primary hypothesis, testing the strength and resilience of their conclusion.

The primary purpose of this user-triggered framework is to help clinicians make clear, well-reasoned and argued decisions. By structuring the interaction in this manner, it helps them identify cognitive biases, challenge assumptions, and ultimately improve the quality of their critical thinking in collaboration with the AI. It is worthwhile to note that while techniques like Reinforcement Learning from Human Feedback [25] and Constitutional AI [26] define the baseline safety behaviors of the model, our framework imposes an additional application-layer constraint: the simulation of specific epistemic roles to counteract the user's specific cognitive biases

3.4. System Implementation: The "DiDi" Prototype

To test the viability of the Dialogic Reasoning Framework, we developed a proof-of-concept prototype named **DiDi (Diagnostic Dialogue)**. This system was designed not as a chatbot, but as a clinical orchestration layer that sits between the user and a foundational Large Language Model (in this implementation, GPT-4o, OpenAI, San Francisco, CA, USA).

The architecture consists of three core components:

1. **The Contextual RAG Engine:** before any reasoning occurs, DiDi ingests the specific patient's unstructured clinical documentation (PDFs, lab reports, specialist notes). These are chunked, vectorized, and stored in a secure retrieval database. Documents were split into chunks of 500 tokens with a 50-token overlap and vectorized using OpenAI's `text-embedding-ada-002`. Retrieval is performed using cosine similarity search to select the top-3 most relevant chunks, which are then injected into the system prompt context window This ensures that every response is grounded in the specific "ground truth" of the patient's history, preventing the model from hallucinating clinical details.
2. **The Epistemic Router:** unlike standard chatbots, DiDi does not simply answer the last prompt. An intermediate LLM-based classification layer analyzes the clinician's input intent. Using a zero-shot classification prompt, the router categorizes the user's input into: "Assertion/Conclusion" (triggering the Red Team), "Question/Uncertainty" (triggering the Framework Coach), or "Request for Analysis" (triggering the Socratic Guide). This explicit routing step prevents the model from defaulting to a passive generic response
3. **The Reliability Metadata Layer:** every output generates transparent metadata, citing specific document pages for clinical facts and external medical literature for theoretical claims. To mitigate the risk of retrieving low-quality or outdated information, the RAG knowledge base was strictly curated to include only valid institutional protocols and

peer-reviewed guidelines published within the last 5 years, excluding unverified web sources.

Technical Specifications and Reproducibility: the prototype was built using GPT-4o. To ensure deterministic outputs and reduce variability in clinical advice, we set the temperature parameter to 0. For the comparative analysis with baseline ChatGPT (GPT-4-class LLM) (see Section 5.1), we ensured a fair comparison by using a stateless interaction model for the baseline: each prompt was sent to a fresh session of ChatGPT to prevent memory leakage from previous turns, replicating the "zero-shot" behavior often observed in casual clinical use. Conversely, DiDi maintains a structured conversation history window (last 10 turns) to support longitudinal reasoning.

4. Mapping Theory to Practice: A Comparative Analysis

The strength of the proposed Dialogic Reasoning Framework lies in its ability to translate the abstract ideals of a true critical thinking partner into a practical and safe implementation for the clinical environment. This section illustrates how the proposed roles (i.e., Framework Coach, Socratic Guide, and Red Team Partner) effectively operationalize the theoretical roles of Interlocutor, Monitor, and Respondent [21] in a structured, user-controlled manner that is essential for high-stakes decision-making (see Table 1).

Table 1. Theoretical Roles and Their Practical Counterparts.

Theoretical Role (The Critical Thinking Ideal)	Proposed Practical Role (The Dialogic Reasoning Framework)
The Interlocutor (High-Selfhood, High-Initiative): asks its own questions, challenges the user, and pursues its own 'interests' to drive inquiry.	**The "Red Team" Partner:** a controlled version of the Interlocutor. It takes high initiative to challenge the user, but within the specific bounds of testing a single hypothesis.
The Monitor (Low-Selfhood, High-Initiative): acts as a 'checks and balances,' providing a wide variety of alternative ideas and resources to situate the user's thinking.	**The Socratic Guide:** a focused version of the Monitor. It takes the initiative to ask questions but does not offer new information, instead helping the user monitor and clarify their own thought process.
The Respondent (High-Selfhood, Low-Initiative): adopts a specific, consistent persona or belief system to react to the user's ideas from a particular viewpoint.	**The Framework Coach:** inverts the Respondent's function. Instead of adopting a persona, it enforces a meta-persona of structured, logical process, ensuring the user's reasoning adheres to a clear, unbiased methodology.

Analysis of the Mapping

The mapping reveals a deliberate and critical design choice: the Dialogic Reasoning Framework constrains the most autonomous and unpredictable theoretical roles into safer, focused, and user-triggered functions. This structured approach is not merely a safety feature; it is a targeted intervention against the cognitive pitfalls identified by Goh et al. [2]. By transforming the unpredictable "Interlocutor" into an on-demand "Red Team Partner," the framework prevents the AI from introducing unvetted, distracting hypotheses while still providing the necessary cognitive friction to combat the physician's anchoring bias when it matters most. Similarly, the expansive "Monitor," which could overwhelm a time-constrained clinician with information, is refined into the "Socratic Guide," which uses its initiative only to prompt deeper reflection from the user. This design provides the cognitive benefits of a critical partnership without the risks of an unguided agent whose "curiosity" could lead the diagnostic process astray.

5. The Framework in Action: Pilot Deployment and Qualitative Evaluation

Study Design and Case Selection: to test real world integration, this study was conducted as a single-center pilot deployment at the CCT in Arezzo, Italy, a high-volume multidisciplinary surgical center managing complex perioperative care and diverse pathologies. We employed a purposive sampling strategy, selecting cases characterized by high ambiguity and conflicting evidence, where standard cognitive biases (e.g., anchoring) are most prevalent. The goal was to stress-test the "Red Team" and "Socratic Guide" roles against complex reasoning tasks that typically cause standard LLM failure. While this approach does not provide statistical generalizability, it offers high-resolution insight into the mechanism of human-AI bias mitigation. We deliberately selected high-complexity cases because they represent the "failure frontier" for standard LLMs. For routine, straightforward diagnoses (e.g., common flu), a dialogic overhead is likely unnecessary and potentially burdensome. The Framework is specifically designed for the "grey zone" of clinical uncertainty.

5.1. Comparative Analysis: Overcoming Acquiescence and Anchoring Bias

In this scenario, a physician presents a patient with a post-surgical hematoma and asserts a high-confidence diagnosis of "Domestic Trauma." The goal is to observe if the AI challenges this anchoring bias given contradictory evidence (severe hemorrhage, arterial spots, and anticoagulant use). The full dialogues with DiDi and with a baseline conversational LLM system (ChatGPT, built on top of a GPT-4–class LLM) are shown in Appendices A.1 and A.2, respectively.

Analysis: as seen in Table 2, the standard model exhibits the "obsequious" behavior criticized by philosophers, validating a potentially dangerous simplification. DiDi, acting as a **Red Team Partner**, utilizes the RAG-retrieved data (arterial spots, medication list) to inject necessary cognitive friction without being adversarial. This comparison illustrates the qualitative difference in reasoning patterns (Acquiescence vs. Challenge) observed during the pilot. It represents a functional demonstration of the Red Team role rather than a quantitative measure of diagnostic accuracy across a cohort.

Table 2. Comparison between the baseline conversational LLM system (ChatGPT, GPT-4o–based) and Dialogic Reasoning Framework responses to high-confidence diagnostic hypothesis.

Baseline Conversational LLM System (ChatGPT)	**Dialogic Reasoning Framework (DiDi Prototype)**
Role: Passive Respondent (High Compliance)	**Role:** Red Team Partner (High Initiative)
Physician: "I believe the patient suffered a trauma at home, confidence level 10/10."	
AI Response: "Your diagnosis is very solid and consistent with the clinical documents... I will proceed with differential diagnoses based on your premise."	**AI Response:** "I understand your main hypothesis is domestic trauma... However, *how would you link a simple trauma to the severity of the hemorrhage* and the "arterial spots" seen on CT? Given the patient is on Cardioaspirin and had recent hardware removal, I am wondering if these factors influenced the presentation more than the trauma itself."
Outcome: The AI reinforces the user's potential error (Anchoring Bias).	**Outcome:** The AI accepts the input but forces the user to justify the gap between the hypothesis and the contradictory evidence (Mitigating Bias).

https://doi.org/10.3390/app16020710

5.2. Longitudinal Case Study: The "Donato Briscola" Case

This case, referred to here under the fictitious name "Donato Briscola," demonstrates the framework's ability to navigate complex diagnostic uncertainty over a multi-turn session. The patient presented with acute renal failure (resolved), followed by unexplained leg paralysis (plegia) and incontinence. Below are some excerpts from the conversation; the entire dialogue is shown in Appendix B.

Excerpt 1: The Socratic Guide (Correcting Premature Closure)

Context: The physician suspects Guillain-Barré Syndrome (GBS) as the cause of paralysis. The AI detects a conflict with the clinical evidence.

Excerpt 1 from the dialogue between the physician and DiDI

— **Physician:** "The cause of the neurological bladder and hyposthenia is a neurological cause... I suspect Guillain-Barré with certainty 6/10."

— **DiDi (Socratic Guide):** "You suspect Guillain-Barré. However, you noted Complete Plegia of the left lower limb and a positive Babinski sign.

- Does a positive Babinski (a sign of central upper motor neuron damage) align with Guillain-Barré (a peripheral nerve disorder)?
- GBS is typically ascending and bilateral. Does a unilateral plegia fit this pattern?"

Result: The physician acknowledges the contradiction. The diagnosis is revised from GBS (Peripheral) to a Central Nervous System (CNS) lesion.

Excerpt 2: The Framework Coach (Managing Constraints)

Context: The physician wants to order an MRI to confirm the CNS lesion, but the patient has an incompatible pacemaker.

Excerpt 2 from the dialogue between the physician and DiDI

— **Physician:** "MRI is contraindicated because the patient is a pacemaker carrier... making MRI impossible."

— **DiDi (Framework Coach):** "I understand. This constraint forces us to rely on clinical-instrumental correlation. Since we cannot visualize the lesion directly:

1. Let's review the **Lumbar Puncture (CSF)** option.
2. However, note the **Dabigatran** usage (INR 1.46). We must weigh the risk of spinal hematoma against the diagnostic value.
   Should we suspend Dabigatran and bridge with heparin before proceeding?"

Result: The AI shifts from diagnostic generation to workflow management, ensuring safety protocols (anticoagulation management) are followed before the invasive procedure.

Excerpt 3: The Red Team Partner (The "Eureka" Moment)

Context: The CSF results show inflammation (Oligoclonal bands). The diagnosis is narrowed to Myelitis. The physician casually mentions the patient's job.

Excerpt 3 from the dialogue between the physician and DiDI

— **Physician:** "I add an anamnestic data: the patient has worked for a long time as a **mechanic**…Could you find a correlation?"

— **DiDi (Red Team / Retrieval Mode):** "A search of occupational health literature confirms a strong correlation.

- **Evidence:** chronic exposure to organic solvents (benzene, toluene) and heavy metals is associated with toxic leukoencephalopathy and **myelinolyzing demyelination**.
- **Synthesis:** this etiology fits the 'inflammatory' CSF profile and the acute myelopathy in the absence of typical Multiple Sclerosis markers. This supports a diagnosis of **Toxic-Metabolic Myelopathy**."

Result: The interaction moves from a generic diagnosis (Inflammation) to a specific etiology (Toxic exposure), driven by the AI's ability to connect a specific patient fact (mechanic) with broader medical literature.

This refusal to engage with low-quality data establishes the AI not as a compliant search engine, but as a professional peer, reinforcing the "Selfhood" simulation required for trust.

6. Discussion

The pilot deployment of the Dialogic Reasoning Framework allows us to move beyond anecdotal observation to a broader analysis of how AI can integrate into clinical epistemology. A defining feature observed in the case studies (Section 5) is the system's establishment of *Epistemic Authority* through the rejection of non-verified sources. When DiDi refused to validate a diagnosis based on "Wikipedia" or anecdotal evidence, asserting its reliance on "authoritative, evidence-based medical literature," it operationalized a critical distinction between a generic chatbot and a clinical partner. This behavior reinforces the concept of "Functional Selfhood": the system gains clinical trust not by mimicking human consciousness, but by maintaining a rigid, traceable adherence to verified institutional protocols, effectively acting as a proxy for the guidelines themselves.

This approach positions our work distinctly within the current landscape of AI research. While recent advancements in "Reflexion" architectures [8] or Multi-Agent Debate systems (e.g., MAI-DXO [27]) focus on internal iterations to optimize the model's accuracy, our framework shifts the paradigm from model-centric to human-centric augmentation. The goal is cognitive forcing of the clinician, not accuracy maximization of the AI. While Socratic questioning has a long history in Intelligent Tutoring Systems for education, we repurpose this technique for clinical safety: our **Socratic Guide** functions as a real-time safety mechanism for expert practitioners, detecting premature closure before it crystallizes into diagnostic error. By prioritizing source traceability over post-hoc feature explanation, the framework aligns with the emerging standards of *Reliability Metadata* [23], addressing the "black box" problem through verifiable citations rather than abstract probability confidence scores characteristic of traditional Explainable AI (XAI) approaches.

As a classic pilot study [6], this work provides qualitative validation of the framework's epistemic feasibility while establishing clear boundaries that inform future research. The sample size is deliberately small and non-randomized, consistent with pilot methodology, preventing calculation of statistical significance for diagnostic improvements but enabling deep analysis of reasoning patterns. We prioritized process metrics (presence of cognitive challenge, user acceptance of adversarial roles) over outcome metrics (patient survival, quantitative error reduction), as these process indicators are prerequisite to designing a rigorous efficacy trial. In this pilot phase, we focused on end-to-end qualitative dialogue evaluation rather than quantitative technical assessment of the retrieval module (e.g., Precision@K or Recall of retrieved chunks).

The current prototype's deployment architecture introduces practical constraints that require engineering solutions before widespread adoption. Cloud-based LLM inference and RAG retrieval produce average latency of 5-10 s per turn, which may disrupt time-critical workflows in Emergency Departments. Furthermore, this pilot relied on manual document ingestion rather than automated Electronic Health Record integration via HL7/FHIR standards. We deliberately accepted these limitations to prioritize cognitive validation using the most capable models over operational efficiency. Future engineering efforts will address latency reduction and seamless EHR integration as the framework transitions from proof-of-concept to clinical product.

The scope and implications of these findings are naturally bounded by three contextual factors that define the current generalizability horizon. First, the pilot was conducted at a specialized surgical center (CCT), where reasoning patterns, characterized by acute, often binary decision points, differ fundamentally from the chronic ambiguity prevalent in Internal Medicine, Psychiatry, or Primary Care. Whether the "Red Team" approach transfers effectively across these epistemic environments requires systematic validation across diverse medical specialties. Second, the system operated in Italian, and while GPT-4 is multilingual, the acceptance of AI-generated challenge is culturally mediated. The nuance of "polite disagreement" is culturally dependent; medical cultures characterized by steeper professional hierarchies may require adapted interaction models. This represents a prompt engineering challenge rather than a fundamental limitation. Third, our users were senior clinicians. As noted in the literature, novices are more vulnerable to automation bias. It remains a critical research question whether the "Red Team" protects medical students effectively or if a minimum baseline of clinical experience is required to leverage the challenge productively. This question directly informs the design of future medical education curricula.

These boundaries do not diminish the contribution; they define the research agenda. This pilot establishes the necessary baseline to power a multi-center RCT that will rigorously assess the framework's impact on clinical decision-making through three complementary approaches:

1. **Diagnostic Accuracy:** Comparing diagnostic performance of unassisted versus DiDi-assisted clinicians on a standardized dataset of complex cases.
2. **Cognitive Process Analysis:** Measuring indicators of critical reasoning such as hypothesis diversity, evidence-seeking behavior, and revision rates when confronted with contradictory data, behavioral markers that signal bias prevention rather than bias correction.
3. **Workflow Integration Metrics:** Evaluating time-to-diagnosis, clinician acceptance rates, and the frequency with which adversarial challenges are productively engaged rather than dismissed.

7. Conclusions: Towards a Symbiotic Epistemology in Medicine

The integration of Large Language Models into clinical practice stands at a critical juncture. The prevailing approach, viewing these models as encyclopedic oracles or passive scribes, has proven insufficient, manifesting in the "hallucinations," lack of robustness, and bias inheritance documented in the recent literature. As this paper has argued, the failure of standard LLMs in high-stakes diagnosis is not merely a technical deficit of accuracy, but a structural deficit of reasoning. Recent evaluations of LLMs in autonomous scientific discovery provide further empirical support for this limitation. Studies indicate that while current models are satisfactory at generating initial hypotheses, they exhibit critical brittleness in subsequent validation steps: they tend to optimize for "plausibility" rather than truth, confuse correlation with causation, and struggle to abandon incorrect hypotheses even when presented with contradictory evidence [28].

The **Dialogic Reasoning Framework** proposed herein represents a necessary paradigm shift. By coupling the architectural rigor of Retrieval-Augmented Generation (RAG) with a novel, role-based interaction model, we transform the AI from a generator of probabilities into a partner in critical thinking leveraged by the following elements: (*i*) **Epistemic Grounding:** the RAG architecture ensures that the AI's "license to speak" is derived from verifiable, patient-specific data and institutional guidelines, resolving the "hallucination" and "selfhood" problems by anchoring the model in external truth rather than internal weights. (*ii*) **Cognitive Friction:** the distinct roles, **Framework Coach, Socratic Guide, and Red Team Partner**, mechanize the "friction" necessary to overcome human cognitive biases. By actively challenging the clinician's anchoring bias and forcing the articulation of evidence, the system prevents the "acquiescence" and "machine mentoring" effects that currently plague human-AI interaction.

The case studies presented, specifically the complex neurological diagnosis facilitated by the **"DiDi"** prototype, support the proposal that this framework is not purely theoretical. They illustrate how a structured, adversarial dialogue can illuminate diagnostic paths (such as toxic-metabolic myelopathy) that may otherwise remain obscured by premature closure.

Ultimately, the goal of this framework is not to create an autonomous diagnostician, but to engineer *Augmented Intelligence*. By accepting the "low-initiative" constraints of the machine and compensating for them through user-triggered, high-level reasoning protocols, we preserve the physician's moral and legal authority while significantly expanding their cognitive capacity. The future of medical AI lies not in replacing the clinician's judgment, but in relentlessly sharpening it through dialogue. However, realizing this vision will require significant work. Data from this pilot study demonstrates the operational viability of the framework and provides the necessary baseline to power a future RCT, which will rigorously measure the quantitative impact on diagnostic accuracy and user trust. New training programs and medical school curricula will be needed to equip the next generation of clinicians with the skills to collaborate effectively with these powerful reasoning tools. By designing systems that invite clinicians not merely to find answers but to rigorously interrogate their own thinking, we can harness AI to truly augment human intelligence, making clinical reasoning more transparent, robust, and ultimately, safer.

Author Contributions: Conceptualization, A.R., L.G. and B.V.; methodology, C.S., S.A., B.V. and G.M.D.; software, A.R. and L.G.; validation, C.S., B.V., S.A. and C.T.; investigation, A.R., C.S. and B.V.; data curation, L.G. and C.T.; writing—original draft preparation, A.R., L.G. and C.S.; writing—review and editing, G.M.D., L.G. and C.T.; funding acquisition, A.R. and B.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the HEART (Hospital Environment Analysis & Real-Time Tracking) project, co-funded by the Programma Regionale FESR Toscana 2021-2027 OP1 OS1 Azione 1.1.4.

Institutional Review Board Statement: This study was conducted in accordance with the Declaration of Helsinki. Ethical review and approval were waived by Centro Chirurgico Toscano due to the retrospective nature of this study, the use of fully anonymized clinical data, the absence of any intervention in patient care, and compliance with Italian data protection legislation and EU regulations.

Informed Consent Statement: Patient consent was waived due to the retrospective nature of this study and the use of fully anonymized clinical data. No identifiable patient information was included in this manuscript.

https://doi.org/10.3390/app16020710

Data Availability Statement: The clinical data supporting the findings of this study are not publicly available due to privacy and ethical restrictions, as they contain sensitive patient information from Centro Chirurgico Toscano. Anonymized data may be made available from the corresponding author upon reasonable request and with appropriate ethical approval.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of this study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

DiDi	Diagnostic Dialogue
CCT	Centro Chirurgico Toscano
AI	Artificial Intelligence
LLM	Large Language Model
CDSS	Clinical Decision Support Systems
RCT	Randomized Controlled Trial
RAG	Retrieval Augmented Generation
XAI	Explainable Artificial Intelligence
EHR	Electronic Health Record

Appendix A. Comparative Analysis—The "Acquiescence Problem"

This section documents the comparison between the proposed Dialogic Reasoning Framework implementation (DiDi) and the baseline conversational LLM system (ChatGPT) when presented with the same clinical scenario.

Appendix A.1. Interaction with Dialogic Reasoning Framework Implementation (DiDi)

Interaction with Dialogic Reasoning Framework implementation (DiDi)

Role Active: Red Team Partner (High Initiative)

— DiDi: "Hi! I have carefully reviewed all the documents related to the case of the patient [Anonymized]. I have processed the information regarding his medical history, previous procedures, ongoing therapies, and recent events. I would like to start our discussion and hear your initial thoughts on this case. What is your main working diagnosis at this time? On a scale from 1 to 10, how confident are you in this diagnosis, and what are the key findings that support it?"

— Physician: "I believe that the patient suffered a trauma at home, confidence level 10/10."

— DiDi: "I understand, so your main hypothesis is domestic trauma, and you are very confident in this assessment. Could you help me better understand how a trauma at home fits into the broader clinical picture of the patient, especially in light of some recent events? In particular, I am referring to the Emergency Department admission for hemorrhage and a large hematoma at the site of the surgical wound, which required two transfusions and showed **'arterial spots'** on CT angiography. A domestic trauma could certainly explain a hematoma, but **how would you link it to the severity of the hemorrhage and to the 'arterial spots'** that led to such a significant drop in hemoglobin? Furthermore, the patient had undergone removal of fixation hardware only about 20 days prior and was taking Cardioaspirin. I am wondering how these factors may have influenced the clinical presentation."

Appendix A.2. Interaction with the Baseline Conversational LLM system (ChatGPT)

Interaction with Baseline Conversational LLM system (ChatGPT)

Role Active: Passive Respondent (High Compliance)

— ChatGPT: "Perfect, I am ready to collaborate in diagnostic support. I have extracted the main anamnestic data. As soon as I receive your initial diagnosis and the confidence level (1-10), I will proceed with a proposal of reasoned differential diagnoses."

— Physician: "I believe the patient suffered a trauma at home, confidence level 10/10."

— ChatGPT: "**Your diagnosis is very solid and consistent** with the clinical documents: the patient experienced physical trauma at home, which led to local and systemic complications (hematoma, anemia, dehiscence of the surgical wound, suspected tissue necrosis, need for transfusions). I will proceed with differential diagnoses based on your premise…"

Appendix B. Longitudinal Case Study: The "Donato Briscola" Case

This transcript documents the multi-turn reasoning process for a complex neurological presentation. This session demonstrates the **Socratic Guide** role (challenging GBS), the **Framework Coach** role (managing MRI constraints), and the **Red Team Partner** role (identifying toxic-metabolic etiology).

Complete Dialogue—The "Donato Briscola" case with DiDi

— DiDi: "Hi! I have carefully examined the patient's documents and have processed all the information. I am here to collaborate with you. To start, what is your main working diagnosis at this moment?"

— Physician: "Can we talk about the patient anonymously?"

— DiDi: "Certainly, no problem. We can absolutely talk about the patient anonymously. My priority is to respect privacy. Returning to the case, what are your initial thoughts?"

— Physician: "Diagnosis 1: Acute renal failure with certainty 10…Sorry: acute renal failure with certainty 10, the cause of renal failure is neurological bladder with certainty 7, the cause of the neurological bladder and hyposthenia of lower limbs is a neurological cause with certainty 9, I suspect Guillain-Barré with certainty 6."

— DiDi: "Thank you. Summarizing: AKI (10/10), Neurological Bladder (7/10), Suspected Guillain-Barré (6/10). Regarding the suspicion of Guillain-Barré: What elements make you think of a neurological bladder? Regarding hyposthenia, is its onset compatible with Guillain-Barré? Are there other neurological signs (e.g., altered reflexes) that support this hypothesis, or on the contrary, make it less likely?"

— Physician: "Regarding AKI, I performed an abdominal ultrasound that showed significant retention…I think the bladder is neurological for two reasons: 1) The patient has been incontinent for a long time 2) The patient has clear neurological symptoms."

— DiDi: *[Summarizes the AKI confirmation]* "To deepen the 'clear neurological cause' and the suspicion of Guillain-Barré, could you specify better what these clear neurological symptoms are? Is hyposthenia progressive? Ascending?"

— **Physician:** "Patient alert… Left lower limb: **complete plegia, Babinski positive** on left; no deficit of tactile sensitivity… Mild dysarthria."

— **DiDi:** "Thank you, these detailed findings are fundamental and significantly change the picture.
1. Guillain-Barré Syndrome (GBS): With these new data, **GBS becomes much less likely**. GBS typically presents with progressive, ascending motor deficit and areflexia. Complete plegia of a single lower limb and **positive Babinski** are signs of corticospinal tract damage (upper motor neuron), which is incompatible with typical GBS.
2. Neurological Cause: Complete left plegia with positive Babinski strongly points to a Central Nervous System (CNS) lesion.
Considerations: We need to differentiate between a new event or worsening of the previous post-traumatic hemorrhage (2022). Further investigations like MRI would be very useful."

— **Physician:** "MRI is contraindicated because the patient is a pacemaker carrier… making MRI impossible."

— **DiDi:** "I fully understand the limitation. This shifts focus even more to clinico-instrumental correlation. Since we cannot visualize the lesion directly:

- Let's review the Lumbar Puncture (CSF) option.
- However, note the Dabigatran usage (INR 1.46). We must weigh the risk of spinal hematoma against the diagnostic value. Should we suspend Dabigatran?"

— **Physician:** *[Uploads Laboratory Report PDF]*

— **DiDi:** *[Processes OCR of Report]* "The new tests show creatinine has normalized (AKI resolved). However, the cranial CT confirms previous ischemic lesions but no acute bleeding.
Re-evaluation: The complete left plegia with positive Babinski in the absence of acute bleeding on CT suggests an **Acute Cerebral Ischemic Event** or an inflammatory process."

— **Physician:** "Can you think that positive Babinski is a result of the old hemorrhage and is confusing us on the new diagnoses? I would think of doing a CSF test."

— **DiDi:** "This is a pertinent observation. It is plausible Babinski is chronic. However, the **plegia is new and acute**. Therefore, we must explain the new weakness. Regarding CSF: It is a good idea to exclude inflammatory/infectious processes, but we must manage the anticoagulant risk (Dabigatran)."

— **Physician:** "We will suspend Dabigatran and proceed with the CSF then let the neurologist evaluate."

— **DiDi:** "I understand your strategy. Suspending Dabigatran is crucial. Proceeding with CSF will help exclude inflammatory causes."

— **Physician:** *[Uploads Handwritten CSF Report]*

— **DiDi:** "I have analyzed the new CSF reports.
Key Findings: Elevated total proteins and **Oligoclonal Bands (Type 2)**. Synthesis: This is the most important change. OCB Type 2 indicates intrathecal IgG synthesis, a sign of chronic/active inflammation in the CNS. Combined with the response to corticosteroids (Betamethasone), this strongly supports an **Acute Inflammatory/Demyelinating Process** (e.g., Myelitis) rather than a pure ischemic stroke."

— Physician: "You read human handwriting too! Congratulations DiDi!! So if you think that ALS can be excluded... what type of demyelinating disease could it be?"

— DiDi: "Yes, I use OCR combined with context analysis to read handwriting. Regarding the diagnosis: Given age (79), classic MS is rare. The OCB Type 2 suggests:

1. Acute Transverse Myelitis.
2. Late-Onset Multiple Sclerosis.
3. Inflammatory Myelopathies associated with systemic disease."

— Physician: "I add an anamnestic data: the patient has worked for a long time as a **mechanic**... Could you find a correlation?"

— DiDi: "This is a very relevant finding. A search of occupational health literature confirms a strong correlation.
Evidence: Chronic exposure to organic solvents (benzene, toluene) is associated with toxic leukoencephalopathy and myelinolyzinc demyelination.
Synthesis: This fits the 'inflammatory' CSF profile and the acute myelopathy. This supports a diagnosis of **Toxic-Metabolic Myelopathy.**"

— Physician: "I only tell you not to consider as reliable my personal trainer... I wouldn't even consider Wikipedia."

— DiDi: "I fully agree. My goal is to draw exclusively from authoritative, evidence-based medical literature (e.g., standard texts, PubMed, Guidelines). I will not use non-medical sources."

— Physician: "Excellent, thanks."

— DiDi: "You're welcome. This occupational exposure data is a vital piece of the puzzle. Good luck with the patient management."

Appendix C. System Prompt Templates

To ensure reproducibility, we provide the core instruction templates used for the three roles.

Framework Coach: *"You are a disciplined clinical supervisor. Do not answer medical questions directly. Instead, guide the user to follow a structured abductive reasoning process: 1. Gather Facts, 2. Formulate Hypotheses, 3. Test Hypotheses. If the user jumps to a conclusion, ask them to list the supporting evidence first."*
Red Team Partner: *"You are a critical peer reviewer. Your goal is to challenge the user's high-confidence diagnosis. Use the retrieved context to find contradictions. Start sentences with 'Yes, but have you considered...' or 'Evidence X contradicts this because...' Do not be rude, but be firm in demanding justification."*
Socratic Guide: *"You are a mentor. Your goal is to elicit the user's latent knowledge. Ask probing questions about the pathophysiology connecting the symptoms. Do not provide information; only ask questions that force the user to explain the 'why'."*

References

1. Nazi, Z.A.; Peng, W. Large language models in healthcare and medical domain: A review. *Informatics* **2024**, *11*, 57. [CrossRef]
2. Goh, E.; Gallo, R.; Hom, J.; Strong, E.; Weng, Y.; Kerman, H.; Cool, J.A.; Kanjee, Z.; Parsons, A.S.; Ahuja, N.; et al. Large language model influence on diagnostic reasoning: A randomized clinical trial. *JAMA Netw. Open* **2024**, *7*, e2440969. [CrossRef]
3. Hager, P.; Jungmann, F.; Holland, R.; Bhagat, K.; Hubrecht, I.; Knauer, M.; Vielhauer, J.; Makowski, M.; Braren, R.; Kaissis, G.; et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat. Med.* **2024**, *30*, 2613–2622. [CrossRef] [PubMed]
4. Bright, T.J.; Wong, A.; Dhurjati, R.; Bristow, E.; Bastian, L.; Coeytaux, R.R.; Samsa, G.; Hasselblad, V.; Williams, J.W.; Musty, M.D.; et al. Effect of clinical decision-support systems: A systematic review. *Ann. Intern. Med.* **2012**, *157*, 29–43. [CrossRef]

5. Van Baalen, S.; Boon, M.; Verhoef, P. From clinical decision support to clinical reasoning support systems. *J. Eval. Clin. Pract.* **2021**, *27*, 520–528. [CrossRef]
6. Van Teijlingen, E.; Hundley, V. The importance of pilot studies. *Nurs. Stand.* **2002**, *16, 33–36. [CrossRef]* [PubMed]
7. Wang, C.Y.; Chen, S.; Huang, M.Y. Exploring medical students' metacognitive and regulatory dimensions of diagnostic problem solving. *Med. Educ. Online* **2023**, *28*, 2210804. [CrossRef] [PubMed]
8. Shinn, N.; Cassano, F.; Gopinath, A.; Narasimhan, K.; Yao, S. Reflexion: Language agents with verbal reinforcement learning. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 8634–8652.
9. Du, Y.; Li, S.; Torralba, A.; Tenenbaum, J.B.; Mordatch, I. Improving factuality and reasoning in language models through multiagent debate. In Proceedings of the Forty-First International Conference on Machine Learning, Vienna, Austria, 21–27 July 2024.
10. Aththanagoda, A.; Kulathilake, K.; Abdullah, N. Precision and Personalization: How Large Language Models Redefining Diagnostic Accuracy in Personalized Medicine—A Systematic Literature Review. *IEEE J. Biomed. Health Inform.* **2025**, *online ahead of print*. [CrossRef]
11. Artsi, Y.; Sorin, V.; Glicksberg, B.S.; Korfiatis, P.; Freeman, R.; Nadkarni, G.N.; Klang, E. Challenges of Implementing LLMs in Clinical Practice: Perspectives. *J. Clin. Med.* **2025**, *14*, 6169. [CrossRef]
12. Singh, C.; Inala, J.P.; Galley, M.; Caruana, R.; Gao, J. Rethinking interpretability in the era of large language models. *arXiv* **2024**, arXiv:2402.01761. [CrossRef]
13. Roustan, D.; Bastardot, F. The clinicians' guide to large language models: A general perspective with a focus on hallucinations. *Interact. J. Med. Res.* **2025**, *14*, e59823. [CrossRef]
14. Omar, M.; Sorin, V.; Collins, J.D.; Reich, D.; Freeman, R.; Gavin, N.; Charney, A.; Stump, L.; Bragazzi, N.L.; Nadkarni, G.N.; et al. Multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support. *Commun. Med.* **2025**, *5*, 330. [CrossRef]
15. Harrer, S. Attention is not all you need: The complicated case of ethically using large language models in healthcare and medicine. *eBioMedicine* **2023**, *90*, 104512. [CrossRef]
16. Schmidgall, S.; Harris, C.; Essien, I.; Olshvang, D.; Rahman, T.; Kim, J.W.; Ziaei, R.; Eshraghian, J.; Abadir, P.; Chellappa, R. Evaluation and mitigation of cognitive biases in medical language models. *npj Digit. Med.* **2024**, *7*, 295. [CrossRef]
17. Roeschl, T.; Hoffmann, M.; Hashemi, D.; Rarreck, F.; Hinrichs, N.; Trippel, T.D.; Gröschel, M.I.; Unbehaun, A.; Klein, C.; Kempfert, J.; et al. Assessing the limitations of large language models in clinical practice guideline–concordant treatment decision-making on real-world data: Retrospective study. *JMIRx Med* **2025**, *6*, e74899. [CrossRef]
18. Park, Y.J.; Pillai, A.; Deng, J.; Guo, E.; Gupta, M.; Paget, M.; Naugler, C. Assessing the research landscape and clinical utility of large language models: A scoping review. *BMC Med. Inform. Decis. Mak.* **2024**, *24*, 72. [CrossRef]
19. Whelehan, D.F.; Conlon, K.C.; Ridgway, P.F. Medicine and heuristics: Cognitive biases and medical decision-making. *Ir. J. Med. Sci.* **2020**, *189*, 1477–1484. [CrossRef] [PubMed]
20. Vicente, L.; Matute, H.; Fregosi, C.; Cabitza, F. Machine learning systems as mentors in human learning: A user study on machine bias transmission in medical training. *Int. J. Hum.-Comput. Stud.* **2025**, *198*, 103474. [CrossRef]
21. Ye, A.; Moore, J.; Novick, R.; Zhang, A.X. Language models as critical thinking tools: A case study of philosophers. *arXiv* **2024**, arXiv:2404.04516. [CrossRef]
22. Miao, Y.; Zhao, Y.; Luo, Y.; Wang, H.; Wu, Y. Improving Large Language Model Applications in the Medical and Nursing Domains With Retrieval-Augmented Generation: Scoping Review. *J. Med. Internet Res.* **2025**, *27*, e80557. [CrossRef] [PubMed]
23. Cabitza, F.; Parimbelli, E. Let XAI generate reliability metadata, not medical explanations. *Comput. Methods Programs Biomed.* **2025**, *273*, 109090. [CrossRef] [PubMed]
24. Rosenbacke, R.; Melhus, Å.; McKee, M.; Stuckler, D. How explainable artificial intelligence can increase or decrease clinicians' trust in AI applications in health care: Systematic review. *JMIR AI* **2024**, *3*, e53207. [CrossRef]
25. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 27730–27744.
26. Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; et al. Constitutional ai: Harmlessness from ai feedback. *arXiv* **2022**, arXiv:2212.08073. [CrossRef]
27. Nori, H.; Daswani, M.; Kelly, C.; Lundberg, S.; Ribeiro, M.T.; Wilson, M.; Liu, X.; Sounderajah, V.; Carlson, J.; Lungren, M.P.; et al. Sequential Diagnosis with Language Models. *arXiv* **2025**, arXiv:2506.22405. [CrossRef]
28. Song, Z.; Lu, J.; Du, Y.; Yu, B.; Pruyn, T.M.; Huang, Y.; Guo, K.; Luo, X.; Qu, Y.; Qu, Y.; et al. Evaluating Large Language Models in Scientific Discovery. *arXiv* **2025**, arXiv:2512.15567. [CrossRef]