

This manuscript proposes a Dialogic Reasoning Framework to address limitations of Large Language Models (LLMs) in clinical decision-making. The authors identify an "acquiescence problem" whereby current LLMs passively confirm clinicians' hypotheses rather than challenging cognitive biases such as anchoring and premature closure. Drawing on empirical evidence showing minimal physician performance improvement with standard LLM assistance (Section 2.1), the authors attribute these failures to the absence of "selfhood" (persistent judgment) and "initiative" (proactive questioning) in current systems (Section 2.2). The proposed framework operationalizes three user-controlled roles—Framework Coach, Socratic Guide, and Red Team Partner—built upon Retrieval-Augmented Generation (RAG) architecture for factual grounding (Section 3). A proof-of-concept implementation ("DiDi") is demonstrated through comparative case studies at Centro Chirurgico Toscano, illustrating how dialogic interactions overcome acquiescence observed in standard LLM responses (Sections 4–5, Appendices A–B).

## Weaknesses

---

### Q1: Insufficient Empirical Validation of Core Claims

The manuscript's central thesis—that the Dialogic Reasoning Framework mitigates cognitive bias and improves diagnostic accuracy—rests almost entirely on anecdotal case studies rather than systematic evaluation.

- **Example 1 (Single-site, non-randomized deployment):** The DiDi prototype was deployed at one institution (Centro Chirurgico Toscano) without description of patient selection criteria, sample size, or comparison group methodology (Section 5, lines 406–416). The authors acknowledge "rigorous clinical validation through randomized controlled trials remains necessary" (lines 16–17) but provide no pilot quantitative metrics (e.g., diagnostic accuracy rates, time-to-diagnosis, inter-rater reliability) that would support feasibility claims.
- **Example 2 (Cherry-picked case presentations):** The two case studies (Sections 5.1–5.2, Appendices A–B) were explicitly "selected to illustrate" specific dimensions (lines 414–416) rather than drawn from a consecutive or randomized sample. No information is provided about cases where the framework failed, produced ambiguous guidance, or was rejected by clinicians, introducing severe selection bias.
- **Example 3 (Absence of comparative performance data):** Table 2 (lines 423–435) presents qualitative dialogue excerpts but no quantitative comparison of diagnostic outcomes. The manuscript cites Goh et al.'s finding that GPT-4 alone achieved 90% accuracy versus 76% with physician assistance (lines 150–152) but provides no analogous metrics for DiDi-assisted versus unassisted diagnosis.
- **Example 4 (No user acceptance or workflow integration data):** Despite claims that the framework addresses "workflow integration" failures (Section 2.1, lines 118–141), no data on clinician adoption rates, time burden, or satisfaction scores are reported. The single user quote rejecting Wikipedia (Appendix B, line 532) is insufficient evidence of epistemic authority acceptance.

---

### Q2: Conceptual Ambiguity in "Selfhood" and "Initiative" Operationalization

The manuscript borrows philosophical constructs from Ye et al. [16] but inadequately reconciles their theoretical definitions with the implemented system's behavior, creating internal contradictions.

- **Example 1 (Contradictory design goals):** The authors state that "strict adherence to concepts like 'selfhood' and 'initiative'... is operationally untenable" due to regulatory/ethical constraints (lines 224–231), yet the framework is explicitly designed to provide "functional benefits of constructive disagreement" (lines 231–232) and "simulate beneficial behaviors like critical questioning" (lines 345–347). The manuscript never clarifies what distinguishes acceptable "simulation" from unacceptable "manifestation" of these properties.
  - **Example 2 (Undefined boundary conditions):** The Red Team Partner is described as maintaining "high initiative to challenge the user" (Table 1, lines 392–407) while remaining "user-triggered" and "controlled" (lines 349–362). The technical implementation details in Section 3.4 do not specify how the "Epistemic Router" (lines 377–381) determines when to activate adversarial roles versus when clinician authority should prevent challenge, leaving the autonomy boundary operationally vague.
  - **Example 3 (Inconsistent role activation logic):** The manuscript states the Red Team Partner activates "when the user issues a command like 'challenge this idea'" (lines 359–361), suggesting explicit user control, but Appendix A.1 shows DiDi initiating Red Team questioning without such a command after the physician's confidence statement (lines 523–527). This discrepancy between described and demonstrated behavior is unexplained.
  - **Example 4 (Lack of failure mode analysis):** The framework aims to prevent "unquantifiable and untraceable bias" from AI "beliefs" (lines 228–229), yet provides no analysis of how the Red Team Partner's "evidence-based alternatives" (lines 9–11) avoid introducing new biases when retrieved evidence is incomplete, contradictory, or reflects systematic literature biases (e.g., publication bias, demographic underrepresentation).
- 

### Q3: Inadequate Treatment of RAG Architecture Limitations

While RAG is positioned as "non-negotiable" (line 267) for epistemic authority, the manuscript glosses over well-documented RAG failure modes that directly threaten the framework's clinical safety claims.

- **Example 1 (Retrieval quality not validated):** Section 3.1 asserts RAG "grounds responses in verifiable data" (lines 269–270) but provides no evaluation of retrieval precision/recall, chunk size optimization, or embedding model performance on medical terminology. The DiDi implementation description (lines 372–375) mentions documents are "chunked, vectorized, and stored" without specifying retrieval algorithms, similarity thresholds, or handling of negated findings (e.g., "no evidence of malignancy" being retrieved for "malignancy" queries).
- **Example 2 (Temporal reasoning gaps acknowledged but unaddressed):** The authors acknowledge RAG systems "often fail to recognize temporal sequences and hierarchical structures critical in medical contexts" (lines 306–308) and that prioritization of older versus newer evidence remains "partially opaque" (lines 308–310). Despite recognizing this as a fundamental limitation, no mitigation strategies are proposed, and the case studies do not demonstrate how DiDi handles conflicting temporal evidence (e.g., outdated guidelines versus recent literature).
- **Example 3 (Source quality control underspecified):** The "Reliability Metadata Layer" (lines 382–384) claims to cite "specific document pages for clinical facts and external medical literature for theoretical claims," but the manuscript provides no taxonomy of acceptable source types, no description of how institutional protocols are validated before ingestion, and no mechanism for handling outdated or retracted literature. The single example of source rejection (Wikipedia, Appendix B) is anecdotal rather than systematic.

- **Example 4 (Contradiction resolution not addressed):** Clinical documents frequently contain contradictory information (e.g., different specialists' assessments, lab values with conflicting interpretations). The manuscript provides no description of how the RAG system prioritizes or synthesizes conflicting retrieved passages, a critical gap given that the Red Team Partner's role depends on presenting "evidence-based alternatives" (lines 9–11).
- 

## Q4: Mathematical and Algorithmic Formulation Absent

The manuscript lacks formal specification of the system's decision-making logic, preventing reproducibility and rigorous analysis of its behavior.

- **Example 1 (No role-selection algorithm):** The "Epistemic Router" (lines 377–381) is described as analyzing "the clinician's input to determine the necessary epistemic role" but no algorithm, decision tree, or classification model is specified. The claim that "high-confidence assertion" triggers Red Team mode (lines 378–379) is not formalized—no threshold for "high confidence" is defined, and no analysis of false-positive/false-negative role assignments is provided.
  - **Example 2 (RAG retrieval scoring undefined):** The manuscript does not specify how retrieved documents are scored for relevance, how many top-k documents are passed to the LLM, or how retrieval scores influence the confidence of generated responses. Standard RAG implementations use cosine similarity or learned reranking; the absence of this specification makes the "traceability" claim (lines 276–277, 382–384) unverifiable.
  - **Example 3 (Prompt engineering not disclosed):** The three roles (Framework Coach, Socratic Guide, Red Team Partner) are described conceptually (Section 3.3) but the actual system prompts that operationalize these roles are not provided. Given that "prompt engineering strategies" are acknowledged as "critically dependent" for LLM performance (lines 63–64), this omission prevents assessment of whether role distinctions are robustly implemented or merely cosmetic variations.
  - **Example 4 (No uncertainty quantification):** The "Reliability Metadata Layer" mentions "confidence scores" and "uncertainty indicators" (lines 313–315) as future enhancements, but the current system provides no formal mechanism for the LLM to express uncertainty or for the clinician to calibrate trust. The case study dialogues (Appendices A–B) show DiDi making definitive statements ("This is a very relevant finding," Appendix B) without quantified confidence bounds.
- 

## Q5: Overstated Claims About Bias Mitigation Without Mechanistic Evidence

The manuscript asserts the framework "systematically mitigates cognitive bias" (lines 11–12) but provides no mechanistic explanation or empirical evidence for how dialogic interaction achieves this effect.

- **Example 1 (Correlation versus causation in case studies):** The comparative analysis (Section 5.1, Table 2) shows DiDi questioning a diagnosis while ChatGPT confirms it, but does not demonstrate that this questioning led to diagnostic accuracy improvement. The physician's response to DiDi's challenge is not reported, leaving unknown whether the "cognitive friction" (lines 472–473) actually changed clinical reasoning or was dismissed.

- **Example 2 (No measurement of bias reduction):** The manuscript cites anchoring bias, confirmation bias, and premature closure (lines 144–145, 163–166) as targets but never operationalizes how these biases would be measured pre/post-intervention. Standard cognitive bias assessment uses metrics like diagnostic revision rates, hypothesis diversity, or evidence-seeking behavior; none are reported.
  - **Example 3 (Potential for new biases unexamined):** While criticizing standard LLMs for "bias transmission" (lines 37–38, 174–192), the manuscript does not analyze whether the Red Team Partner's adversarial stance could introduce oppositional bias (over-challenging correct diagnoses) or whether the Framework Coach's structured process could induce procedural rigidity that delays time-sensitive diagnoses.
  - **Example 4 (Generalization across bias types unsupported):** The case studies focus on anchoring bias (domestic trauma case, Section 5.1) and premature closure (Guillain-Barré case, Section 5.2) but do not address other documented biases like availability heuristic, representativeness heuristic, or affective bias. The claim of "systematic" mitigation (line 12) is not supported by this narrow coverage.
- 

## Q6: Insufficient Discussion of Practical Deployment Barriers

The manuscript underestimates implementation challenges that would prevent real-world adoption of the proposed framework.

- **Example 1 (Computational and latency costs unaddressed):** RAG systems require real-time document retrieval, embedding computation, and multi-turn LLM inference. No analysis of response latency, computational requirements, or cost per query is provided. Given that clinical workflows are time-constrained (emergency settings, surgical decision-making), the feasibility of multi-turn dialogues is questionable but not discussed.
  - **Example 2 (Integration with existing EHR systems):** The manuscript mentions ingesting "unstructured clinical documentation (PDFs, lab reports, specialist notes)" (lines 373–374) but does not address interoperability standards (HL7 FHIR, DICOM), data extraction from proprietary EHR formats, or handling of real-time data updates. The Centro Chirurgico Toscano deployment details are not provided, leaving unknown whether integration challenges were encountered.
  - **Example 3 (Liability and regulatory approval pathways unclear):** The framework is described as a "clinical decision support system" (line 18) which in many jurisdictions requires regulatory approval (FDA Class II/III in US, CE marking in EU). The manuscript does not discuss the approval pathway, liability allocation when AI-generated challenges lead to adverse outcomes, or documentation requirements for medicolegal defense.
  - **Example 4 (Clinician training requirements underspecified):** The authors note "new training programs and medical school curricula will be needed" (lines 489–490) but provide no analysis of the learning curve, the skills required to effectively use adversarial AI partners, or how to prevent over-reliance or under-reliance on system recommendations. The single-site deployment provides no data on training duration or user competency assessment.
  - **Example 5 (Scalability across medical specialties not addressed):** The case studies focus on surgical/neurological cases at one center. No discussion addresses whether the framework generalizes to other specialties (oncology, psychiatry, pediatrics) with different reasoning patterns, evidence bases, and guideline structures, or whether specialty-specific customization would be required.
-

## Q7: Limited Engagement with Alternative Approaches and Related Work

The manuscript positions the Dialogic Reasoning Framework as a novel solution but insufficiently compares it to existing clinical decision support paradigms and recent AI safety research.

- **Example 1 (Sparse comparison to traditional CDSS):** The authors cite Bright et al. [4] on clinical decision support systems (line 542) but do not compare the dialogic approach to established CDSS architectures (rule-based systems, Bayesian networks, case-based reasoning). The claim that LLMs should not be "an alternative to clinical support system tools" (line 41) is not substantiated with comparative analysis of when traditional CDSS might be superior.
  - **Example 2 (Recent LLM safety literature underrepresented):** While the manuscript cites work on hallucination (references 7–9) and bias (references 11, 15), it does not engage with recent literature on constitutional AI, reinforcement learning from human feedback (RLHF) for safety, or multi-agent debate systems that address similar problems through different architectures. The novelty claim would be strengthened by explicit comparison to these alternatives.
  - **Example 3 (Socratic questioning in AI not novel):** The "Socratic Guide" role (lines 356–358) is presented as innovative, but Socratic dialogue systems have been explored in educational AI and intelligent tutoring systems for decades. No citation or comparison to this prior work is provided, making it unclear what is architecturally novel versus a reapplication of known techniques.
  - **Example 4 (Explainable AI (XAI) literature engagement superficial):** The manuscript briefly mentions XAI (lines 312–313) and cites Cabitza & Parimbelli [18] on reliability metadata, but does not engage with the extensive XAI literature on counterfactual explanations, attention visualization, or concept-based explanations that could complement or substitute for the dialogic approach in achieving transparency.
- 

## Q8: Ethical Considerations and Potential Harms Underexplored

The manuscript briefly mentions ethical constraints (lines 226–231, 344–348) but does not provide a comprehensive ethical analysis of the framework's potential harms.

- **Example 1 (Adversarial interaction psychological impact):** The Red Team Partner is designed to challenge clinicians' diagnoses (lines 359–361, Table 2), but no discussion addresses the psychological impact of repeated adversarial interactions on clinician confidence, job satisfaction, or burnout. Research on automation-induced stress is not cited.
- **Example 2 (Equity and access disparities):** The framework requires substantial computational infrastructure (RAG databases, LLM inference) and institutional resources (curated guidelines, EHR integration). No discussion addresses whether this would exacerbate healthcare disparities between well-resourced and under-resourced settings, or how the framework could be adapted for resource-limited environments.
- **Example 3 (Patient autonomy and consent):** The manuscript mentions "Informed Consent Statement: Patient consent was waived" (lines 505–507) for the retrospective study, but does not discuss whether prospective deployment would require patient consent for AI involvement in diagnosis, how patients would be informed about AI's role, or how patient preferences about AI use would be accommodated.

- **Example 4 (Accountability gaps in multi-agent systems):** When the Framework Coach, Socratic Guide, and Red Team Partner provide conflicting guidance, the manuscript does not specify how accountability is allocated if an adverse outcome occurs. The statement that "final moral and legal responsibility... must remain with the human clinician" (lines 69–70) does not address whether this is legally defensible when AI systems actively challenge clinical judgment.
- 

## Q9: Writing Clarity and Organizational Issues

While generally well-written, the manuscript contains organizational and clarity issues that impede comprehension.

- **Example 1 (Inconsistent terminology):** The system is referred to as "Dialogic Reasoning Framework" (title, lines 7–8, 43), "Dialogical AI" (title), and "DiDi" (lines 14, 367) without clear explanation of their relationship until Section 3.4. The distinction between the conceptual framework and the specific implementation should be clarified earlier.
  - **Example 2 (Section 2 structure confusing):** Section 2 is titled "The Challenge: Why Standard LLMs Falter" but contains both Section 2.1 ("Empirical Evidence") and Section 2.2 ("A Deeper Diagnosis"), creating a three-level hierarchy (Section 2 → 2.1 → subsections like "Inaccuracy and Generation of False Content") that is difficult to navigate. Flattening or restructuring would improve readability.
  - **Example 3 (Redundancy between main text and appendices):** The comparative analysis in Section 5.1 (Table 2) largely duplicates content from Appendices A.1 and A.2. Either the table should summarize more concisely or the appendices should provide additional detail not in the main text.
  - **Example 4 (Figure/table absence):** The manuscript contains only two tables (Tables 1–2) and no figures. A system architecture diagram showing the RAG pipeline, Epistemic Router, and role interactions would significantly clarify the technical implementation (Section 3.4). A flowchart of the diagnostic reasoning process would help visualize the Framework Coach's structured approach.
  - **Example 5 (Abbreviation inconsistency):** The abbreviations section (lines 516–518) lists "DiDi" as "Diagnostic Dialogue" but this expansion is not used in the main text. "CCT" is defined in the abbreviations but also spelled out multiple times in the text (lines 407, 505).
- 

## Q10: Generalizability and External Validity Concerns

The single-site, single-specialty deployment limits the generalizability of the findings.

- **Example 1 (Institutional specificity):** The framework was tested at Centro Chirurgico Toscano, described as a "high-volume multidisciplinary surgical center" (lines 408–409). Surgical decision-making has different time pressures, evidence bases, and reasoning patterns than primary care, emergency medicine, or chronic disease management. No analysis addresses transferability.
- **Example 2 (Language and cultural context):** The case study dialogues (Appendices A–B) appear to be translated from Italian (patient name "Donato Briscola," institutional location in Arezzo, Italy). No discussion addresses whether the framework's performance depends on language-specific LLM capabilities, whether medical terminology translation affects retrieval quality, or how cultural differences in clinical communication norms might affect acceptance of adversarial AI.

- **Example 3 (Expertise level of users):** The manuscript describes users as "senior clinicians" (line 411) without specifying their training level, years of experience, or baseline diagnostic accuracy. The framework's effectiveness may differ for novices versus experts (as suggested by Vicente et al.'s finding that lower-baseline students are more vulnerable to AI bias, lines 181–183), but this is not explored.
- **Example 4 (Case complexity selection bias):** The two presented cases involve complex, ambiguous diagnoses (post-surgical hematoma with multiple complicating factors; acute neurological presentation with differential diagnosis evolution). No discussion addresses whether the framework provides value for straightforward cases or whether it would introduce unnecessary complexity and time burden when diagnoses are clear.