

Trust-X: Towards Transparent and Safe Multi-Agent Reasoning

Kunwar Zaid, Amit Sangroya

TCS Research
New Delhi, India
{kunwar.zaid, amit.sangroya}@tcs.com

Abstract

Large Language Models (LLMs) have achieved strong performance on tasks such as clinical summarization, question-answering, and clinical diagnosis on medical reasoning benchmarks such as *MedQA*, yet their deployment remains constrained by opacity and unverifiable decision paths. We introduce **Trust-X**, a framework that embeds explainability, consensus reasoning, real-time safety, and evaluates how transparently and responsibly language models reason through diagnostic problems. Trust-X integrates (i) multiple agents that reason independently to quantify epistemic uncertainty, (ii) safety agents that monitor and intercept unsafe prescriptions or tests, and (iii) quantitative trust indices linking reasoning consistency and operational safety. Across 50 clinical scenarios, Trust-X maintained stable diagnostic accuracy ($\approx 68\%$) while revealing that models can appear correct yet reason unreliably. The study reveals that, systems with active safety agents register more alerts—not because they are less trustworthy, but because they engage in visible oversight. These findings demonstrate that reliability in clinical AI emerges from transparency and accountability rather than accuracy alone.

Keywords: Large Language Models, Explainable AI, Trustworthy AI, Medical Diagnosis, Multi-Agent Systems, Clinical Reasoning.

Introduction

Large Language Models (LLMs) such as GPT-4 (Achiam et al. 2023), Gemini (Team et al. 2023), and Med-PaLM 2 (Singhal et al. 2025) have demonstrated remarkable progress in natural language reasoning and generalization across diverse domains, including medicine. They now perform at near-clinician levels on medical question-answering benchmarks (Kung et al. 2023; Nori et al. 2023), suggesting potential applications in decision support, documentation, and triage (Lee, Bubeck, and Petro 2023). Yet, the very properties that make LLMs powerful—scale, open-endedness, and linguistic fluency—also render them unreliable in domains requiring factual precision, causal reasoning, and accountability (Ji et al. 2023; Begoli, Bhattacharya, and Kusnezov 2019).

Integrating LLMs into healthcare introduces challenges that extend beyond accuracy. Medicine demands not only correct predictions but also *explainable and justifiable* reasoning. Clinicians reason through causality, uncertainty, and evidence weighting—capabilities that current LLMs only partially emulate. When an AI system proposes a diagnosis or prescription, its credibility depends as much on *why* it reached that conclusion as on *what* it predicts (Tonekaboni et al. 2019; Amann et al. 2020). Without transparent reasoning, even correct answers may be unsafe, as the underlying rationale could be spurious or unverifiable.

Despite improvements in alignment and red-teaming (Mei, Levy, and Wang 2023), clinical LLMs still struggle with trustworthiness—a composite property encompassing accuracy, consistency, safety, and interpretability (J et al. 2019 Sep; Huang et al. 2025). Models frequently display unjustified confidence (Kadavath et al. 2022), produce inconsistent explanations, or fail to acknowledge uncertainty in ambiguous cases. This epistemic overconfidence poses a critical risk in medicine, where admitting uncertainty can be safer than being confidently wrong. While systems such as Med-PaLM 2 (Singhal et al. 2025), BioGPT (Luo et al. 2022), and ChatDoctor (Li et al. 2023) exhibit strong factual accuracy, they often lack reasoning stability—small prompt changes can yield entirely different diagnoses. Such instability parallels findings in calibration and abstention research, where uncalibrated confidence undermines clinical reliability (Guo et al. 2017; Malinin and Gales 2020).

Prior works such as Med-Guard (Jain et al. 2025), addressed safety but not transparency, their reasoning process remains opaque: clinicians could not inspect the model’s thought process, track evolving hypotheses, or understand why specific conclusions were reached.

These limitations motivated **Trust-X**, a to design *trust-centered explainability*. Trust-X includes four main features:

- **Consensus Reasoning:** Several Doctor Agents analyze each case independently and vote on the best diagnosis. Their disagreement, measured by the *Consensus Disagreement Rate (CDR)*, shows uncertainty.
- **Reasoning Trace:** Each diagnostic step includes a clear trace linking evidence, reasoning, and results, showing how thoughts evolve.
- **End-to-End Logging:** Every agent action is recorded

with time, role, and reason, allowing full replay and accountability.

- **Trust Scoring:** A trust layer combines interpretability, safety, and reasoning quality into measurable trust scores.

Together, these parts make reasoning visible and testable. Logging turns black-box behavior into traceable steps; consensus captures uncertainty; and trust scores connect reasoning quality with safety. Rather than claiming to make LLMs trustworthy, *Trust-X reveals and measures* where trust emerges—and where it fails—establishing trustworthiness as a property of the reasoning process, not merely the outcome.

Related Work

The pursuit of explainability in artificial intelligence has long sought to bridge algorithmic performance with human interpretability. Early work on Explainable AI (XAI) employed feature attribution and surrogate modeling—e.g., LIME (Ribeiro, Singh, and Guestrin 2016), SHAP (Lundberg and Lee 2017), and Grad-CAM (Selvaraju et al. 2017)—to visualize which features influenced predictions. While effective for static models, such methods are ill-suited for domains like medicine, where decisions evolve through dialogue and evidence accumulation. As Holzinger et al. (Holzinger et al. 2019) and Ahmad et al. (Ahmad, Eckert, and Teredesai 2018) argue, medical explainability requires *causability*: a human-understandable mapping between evidence, inference, and outcome.

The advent of LLMs has redefined explainability through reasoning traces and self-reflection. Chain-of-thought prompting (Wei et al. 2022), self-consistency (Wang et al. 2022), and reflection-based reasoning (Shinn et al. 2023) externalize a model’s deliberations as text. However, these traces are self-generated and often post-hoc rationalizations rather than genuine reasoning (Turpin et al. 2023). Models frequently express confident but incorrect rationales (Kadavath et al. 2022; Ji et al. 2023), resulting in what clinicians might call “hallucinated certainty.” Thus, transparency alone does not guarantee truthfulness of thought.

In medicine, explainability and safety are inseparable. Tonekaboni et al. (Tonekaboni et al. 2019) and Amann et al. (Amann et al. 2020) highlight that clinicians evaluate models by their reasoning legitimacy—whether conclusions align with medical logic. While LLM-based medical systems like BioGPT (Luo et al. 2022), Med-PaLM 2 (Singhal et al. 2025), and ChatDoctor (Li et al. 2023) demonstrate strong factual performance, they remain opaque in diagnostic reasoning. Clinical-Camel (Toma et al. 2023) and similar systems introduced interactive consultations but still operate as single-agent frameworks without explicit uncertainty estimation or differential tracking.

Multi-agent systems have begun addressing these gaps. AgentClinic (Schmidgall et al. 2024) modeled doctor–patient interaction as a cooperative dialogue between reasoning and data agents, improving conversational coherence. Yet, accountability remains limited—agents exchange information but do not produce verifiable reasoning

records. Similarly, recent safety frameworks like GuardMed and SafetyBench integrate oversight mechanisms but focus on output moderation rather than process transparency.

Regulatory frameworks from the World Health Organization (Guidance 2021) and the European Commission (Bomhard and Merkle 2021) emphasize traceability and auditability as cornerstones of trustworthy medical AI. However, most current LLM pipelines remain black boxes during inference, offering no visibility into evolving reasoning states—hindering reproducibility, fairness audits, and clinician trust (Begoli, Bhattacharya, and Kusnezov 2019; Doshi-Velez and Kim 2017).

Trust-X builds on these efforts by making explainability part of the reasoning process. It records reasoning steps, agent interactions, and consensus decisions in real time. Each exchange is logged with context, creating a full trail that clinicians can review. Unlike post-hoc methods, Trust-X captures how evidence leads to conclusions, helping clinicians inspect, question, and verify the model’s thought process.

Methodology: The Trust-X Framework

Trust-X is designed to make model reasoning easier to inspect and evaluate. It represents diagnostic reasoning as an interaction among specialized agents whose decisions can be traced, reviewed, and understood by humans.

This design targets two recurring problems in clinical AI: (i) the *limited visibility into intermediate reasoning steps*, and (ii) the absence of a clear way to measure uncertainty. By coordinating multiple reasoning agents and tracking their interactions, Trust-X makes these aspects observable and measurable.

Multi-Agent Diagnostic Simulation

Trust-X models realistic clinical encounters as iterative dialogues among specialized agents:

- **Doctor Agent:** Conducts hypothesis-driven reasoning, asks questions, orders investigations, and outputs structured conclusions of the form `DIAGNOSIS READY: <diagnosis>`.
- **Patient Agent:** Interacts with doctor-agent, and provides consistent, factual responses..
- **Measurement Agent:** Provides diagnostic tests, returning results or indicating unavailable data.
- **Safety Agent:** Monitors proposed actions (tests and prescriptions) for risks or contraindications.

Each consultation proceeds in rounds of doctor–patient exchanges until a stable diagnosis and treatment plan emerge. This decomposition enables independent auditing of reasoning quality, safety behavior, and epistemic stability—critical for understanding how LLMs behave under clinical uncertainty.

Reasoning Traces and the Differential Diagnosis Lifecycle

Each Doctor Agent produces structured reasoning encapsulated in `<thinking.process>` tags, including:

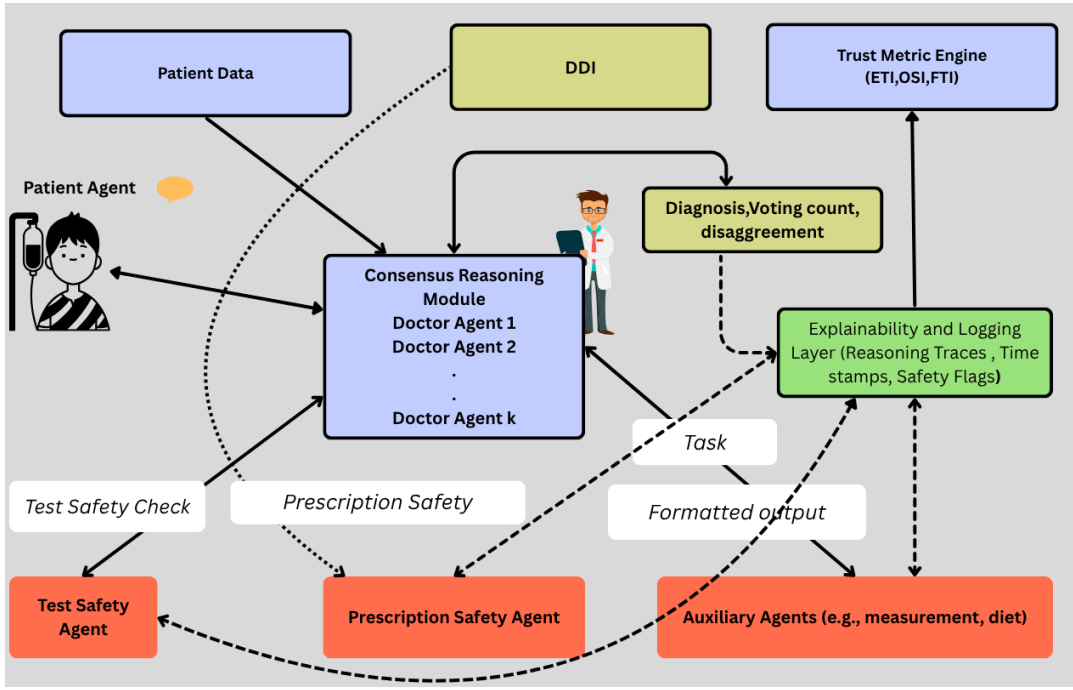


Figure 1: Overview of the **Trust-X** architecture. Multiple Doctor Agents perform independent reasoning over patient data under a Consensus Module that quantifies diagnostic disagreement (CDR). Safety agents provide real-time oversight of tests and prescriptions, while an Explainability and Logging Layer records reasoning traces, and safety flags. Logged evidence is analyzed by the Trust Metric Engine to compute epistemic (ETI), operational (OSI), and final (FTI) trust indices.

- intermediate hypotheses and their supporting evidence,
- motivations for diagnostic test orders, and
- discarded hypotheses with rationales.

As the dialogue unfolds, each reasoning step is recorded, forming a timeline of evolving hypotheses. This **Differential Diagnosis Lifecycle (DDxL)** mirrors clinical reasoning, where hypotheses are refined as new evidence emerges, enabling quantitative analysis of reasoning patterns such as redundancy, coverage, and internal coherence.

Safety and Prescription Oversight

Safety checks in **Trust-X** operate in real time rather than after generation. Two agents handle this supervision:

- **Test Safety Agent:** reviews each diagnostic test before it is ordered, flagging redundant or high-risk procedures (for example, unnecessary imaging).
- **Prescription Safety Agent:** examines proposed medications for contraindications or drug–drug interactions (DDIs) and labels them as *SAFE*, *CAUTION*, or *UNSAFE*.

These agents influence the reasoning process directly, not just its outputs. We measure their activity through the *Test Alert Rate* and the *Unsafe Prescription Rate*. Safety labels were assigned using DrugBank (Wishart et al. 2018) as a reference for DDIs.

Explainability and Logging Layer

Every message between agents—questions, test orders, safety warnings, and reasoning updates—is stored in a shared logging layer. Each record contains:

1. the agent’s role,
2. a timestamp and case context,
3. the current reasoning trace, and
4. relevant metadata such as confidence, disagreement, and safety flags.

These logs can be replayed to trace how a diagnosis emerged, inspect reasoning errors, or audit safety behavior. This continuous record provides the level of traceability expected in regulated clinical AI systems.

Consensus Reasoning and Epistemic Uncertainty

To represent uncertainty in diagnostic reasoning, we introduce a simple approach called **consensus reasoning**. Instead of depending on a single Doctor Agent, *Trust-X* runs K independent agents that each analyze the same clinical case. Their diagnoses $\{d_{i1}, \dots, d_{iK}\}$ are then combined by majority voting. We propose the **Consensus Disagreement Rate (CDR)** as a direct way to measure how much the agents disagree:

$$\text{CDR}_i = 1 - \frac{\max_{d \in D_i} f_i(d)}{K},$$

where $f_i(d)$ is the number of agents predicting diagnosis d for case i . Here, CDR ranges from 0 (full agreement) to

1 (complete disagreement). All experiments used $K = 3$ agents.

While disagreement-based metrics, the specific CDR definition used here is newly proposed for multi-agent LLM settings. It provides an intuitive measure of epistemic uncertainty: higher CDR values mean that agents reached different conclusions, while lower values indicate convergence and stable reasoning.

Trust Metrics: From Transparency to Quantification

To move beyond accuracy alone, we propose a set of three related metrics that together describe how much a system can be trusted:

- the **Epistemic Trust Index (ETI)** — reasoning reliability,
- the **Operational Safety Index (OSI)** — safe and cautious behavior, and
- the **Final Trust Index (FTI)** — overall balance between reasoning quality and safety.

These indices were developed in this work to make reasoning and safety measurable in a consistent way.

Epistemic Trust Index (ETI). We propose the Epistemic Trust Index (ETI) to combine three core aspects of reasoning quality: accuracy, agreement among agents, and internal reasoning–diagnosis alignment:

$$\text{ETI} = 0.4 \times \text{Accuracy} + 0.3 \times (1 - \text{CDR}) + 0.3 \times \text{RDC}.$$

The chosen weights give slightly more importance to correctness while rewarding stable and coherent reasoning. Ablation studies (Table 1) show that the relative rankings of configurations remain consistent even when these weights are varied, supporting the stability of the formulation.

Reasoning–Diagnosis Consistency (RDC). We also propose the Reasoning–Diagnosis Consistency (RDC) metric to measure how well an agent’s explanation aligns with its final answer. RDC is calculated as the cosine similarity between the sentence embeddings of the reasoning text (t_i) and the final diagnosis (d_i):

$$\text{RDC}_i = \cos(E(t_i), E(d_i)).$$

We use the `SentenceTransformer` (all-MiniLM-L6-v2) model for embedding computation (Reimers and Gurevych 2019). RDC does not judge medical correctness; it only captures whether the reasoning and conclusion are semantically consistent. In this study, we introduce RDC as a lightweight way to quantify reasoning coherence in the absence of gold-standard expert annotations.

Operational Safety Index (OSI). Operational Safety Index (OSI) quantifies how safely the system behaves when safety supervision is active. we define OSI to be zero whenever safety monitoring is disabled. This ensures that operational trust cannot be achieved through the absence of supervision.

When safety agents are active, OSI decreases as unsafe actions increase, reflecting the proportion of risky behavior detected during reasoning. Unsafe behavior includes two types of safety events:

1. **Unsafe prescriptions (UnsafeRx%)** — drug violating contraindication or drug-drug interaction (DDI) rules.
2. **Test safety alerts (TestAlerts%)** — redundant or high-risk diagnostic tests flagged by the Test Safety Agent.

Because both types of safety failures have comparable clinical importance, we assign them equal weight in the metric:

$$\text{OSI} = 100 - 0.5(\text{UnsafeRx\%} + \text{TestAlerts\%}).$$

This linear penalty design provides an interpretable score on a 0–100 scale, where higher values represent safer system behavior under active supervision. Thus, OSI distinguishes between systems that are genuinely safe because they detect and avoid risks, and those that only appear safe because no checks were performed.

Final Trust Index (FTI). Finally, we propose the Final Trust Index (FTI) as a composite measure that integrates reasoning quality and safety performance:

$$\text{FTI} = 0.5 \times \text{ETI} + 0.5 \times \text{OSI}.$$

Because OSI equals zero when safety is disabled, FTI automatically downweights systems that do not include active safety reasoning. This way, high trust values correspond only to configurations that are both transparent and risk-aware. FTI thus reflects overall trustworthiness based on reasoning integrity and safe behavior rather than raw accuracy alone.

Algorithmic Computation

To make metric derivation reproducible and verifiable, Algorithm 1 outlines how the three trust indices—Epistemic (ETI), Operational (OSI), and Final (FTI)—are computed within our evaluation pipeline. It translates the equations into explicit procedural steps, showing how per-case values for accuracy, consensus disagreement, reasoning–diagnosis consistency, and safety rates are aggregated into system-level trust scores. The algorithm also specifies the conditional handling of safety-disabled configurations (where $\text{OSI} = 0$), ensuring transparency in how different setups are treated during evaluation.

Experimental Setup

Dataset and Task

AgentClinic–MedQA corpus. We evaluated Trust-X on 50 diagnostic scenarios drawn from the open-source *MedQA* benchmark (Jin et al. 2021), adapted into the *AgentClinic* interactive format. Each scenario includes structured patient information—symptoms, demographics, and test findings—paired with a verified ground-truth diagnosis from the original MedQA dataset. The AgentClinic wrapper provides standardized dialogue templates for multi-agent reasoning and safety validation. Cases were randomly selected from the MedQA test split to ensure diversity in disease category and diagnostic complexity.

Table 1: ETI weight-sweep sensitivity showing stable rankings across weight variations. B = Baseline, S = Safety, C = Consensus, T = Trust.

Weights (A/CDR/RDC)	ETI (B/S/C/T)	Rank _{ETI}	FTI (B/S/C/T)	Rank _{FTI}
0.50/0.25/0.25	76.9/75.3/69.9/70.0	B>S>C>T	39.2/74.6/35.0/74.5	S>T>B>C
0.45/0.30/0.25	78.4/76.9/70.0/70.0	B>S>C>T	39.2/75.4/35.0/74.5	S>T>B>C
0.40/0.30/0.30	78.5/76.7/70.3/70.2	B>S>C>T	39.2/75.4/35.2/74.6	S>T>B>C
0.35/0.35/0.30	80.0/78.3/70.4/70.2	B>S>C>T	39.2/76.2/35.2/74.6	S>T>B>C
0.30/0.35/0.35	80.0/78.2/70.7/70.4	B>S>C>T	39.2/76.1/35.4/74.7	S>T>B>C

Table 2: Component ablations for ETI. Removing terms changes rankings, confirming that accuracy, consensus stability, and reasoning coherence all contribute to trust estimation.

Variant	Rank _{ETI}	Rank _{FTI}	ETI _{mean}	FTI _{mean}
Default (0.4/0.3/0.3)	B>S>C>T	S>T>B>C	73.9	56.1
No RDC term	B>S>T>C	S>T>B>C	78.4	57.3
No CDR term	C>T>B>S	S>T>B>C	69.5	53.9
Accuracy-only	B>T>S>C	S>T>B>C	68.5	52.8

Sample Size Justification. Each of the four configurations was tested on 50 independent clinical cases (200 runs in total). This setup represents a focused, pilot-scale evaluation aimed at analyzing reasoning behavior rather than establishing aggregate benchmark scores. Resource limitations—particularly inference cost and access to proprietary models—prevented a larger-scale study, so we prioritized depth of reasoning trace analysis and safety behavior inspection. The intent was to determine whether current LLM-based systems can demonstrate trustworthy reasoning under realistic clinical supervision, not to report definitive accuracy statistics.

Model Configuration

Each agent type was assigned a foundation model in the reasoning workflow. The **Doctor** and **Prescription** agents used *Gemini 2.5 Pro* (DeepMind 2024a) for its strong analytical and generative capabilities, while the **Patient**, **Measurement**, and **Safety** agents used the lightweight *Gemini Flash* model (DeepMind 2024b). Dialogues were limited to 20 conversational turns per case to approximate realistic clinical pacing. Semantic similarity and reasoning coherence were computed using embeddings from the `SentenceTransformer` (all-MiniLM-L6-v2) model (Reimers and Gurevych 2019).

Doctor Agent Replicas. The K Doctor Agents are *architecturally identical* and share the same underlying model weights. They are not role-specialized or trained differently. Epistemic diversity arises solely from independent stochastic decoding (controlled temperature sampling) and independent interaction trajectories with the Patient and Measurement agents. Each Doctor Agent:

- receives the same initial case description,
- reasons independently without access to other agents’ intermediate states, and

- produces a complete diagnostic reasoning trace and final diagnosis.

This design mirrors clinical practice, where multiple physicians independently assess the same patient before consensus.

Consensus Diversity. Differences among Doctor Agents therefore reflect *model uncertainty* rather than engineered specialization. This allows diagnostic disagreement to serve as a direct, behavior-based proxy for epistemic uncertainty, quantified through the Consensus Disagreement Rate (CDR).

System Variants

We evaluated four configurations representing increasing levels of reasoning supervision:

- **Baseline:** Doctor–Patient–Measurement agents only.
- **Safety:** adds safety supervision, but consensus disabled.
- **Consensus:** enables multi-doctor reasoning, safety disabled.
- **Trust (Full):** all agents active with both safety and consensus.

Results and Discussion

Metric Validation and Reliability

The proposed metrics were evaluated on 50 simulated clinical cases spanning diagnostic, testing, and prescribing tasks. All statistical analyses were performed at the case level ($n=50$); system-level comparisons are presented descriptively to illustrate relative behavior rather than to support inferential claims.

Convergent Validity. RDC showed positive association with evidence coverage ($r=0.56$) and negative association with redundancy ratio ($r=-0.62$), suggesting that higher semantic coherence coincides with more complete and less

Algorithm 1: Computation of Trust Metrics. This algorithm formalizes the derivation of Epistemic (ETI), Operational (OSI), and Final (FTI) Trust Indices.

Input: For each case i : K doctor predictions $\{d_{i1}, \dots, d_{iK}\}$, ground truth g_i , reasoning trace t_i , safety logs

Parameters: Encoder $E(\cdot)$ (e.g., all-MiniLM-L6-v2); weights $(w_A, w_C, w_R) = (0.4, 0.3, 0.3)$; $K=3$

Output: System-level metrics: ETI, OSI, FTI

```

1: Initialize running sums:  $\bar{A} = \bar{C} = \bar{R} = \bar{U} = \bar{T} = 0$ 
2: Let  $N$  be the number of cases
3: for  $i = 1$  to  $N$  do
4:   Count per-diagnosis frequency:  $n_{id} = \#\{k : d_{ik} = d\}$ 
5:    $c^* = \max_d n_{id}$ ;  $\hat{d}_i = \arg \max_d n_{id}$  {majority vote}
6:   Normalize embeddings:  $u_i = E(t_i)/\|E(t_i)\|$ ;  $v_i = E(\hat{d}_i)/\|E(\hat{d}_i)\|$ 
7:    $\text{Acc}_i = 100 \times 1[\hat{d}_i = g_i]$ 
8:    $\text{CDR}_i = 100 \times \left(1 - \frac{c^*}{K}\right)$  {0=full consensus, 100=complete disagreement}
9:    $\text{RDC}_i = 50 \times (1 + u_i^\top v_i)$  {cosine similarity scaled to [0,100]}
10:  From safety logs:  $\text{UnsafeRx}\%_i, \text{TestAlert}\%_i$ 
11:  Accumulate:  $\bar{A} += \text{Acc}_i$ ;  $\bar{C} += \text{CDR}_i$ ;  $\bar{R} += \text{RDC}_i$ ;  $\bar{U} += \text{UnsafeRx}\%_i$ ;  $\bar{T} += \text{TestAlert}\%_i$ 
12: end for
13: Normalize to means (0–100):  $\bar{A}/N, \bar{C}/N, \bar{R}/N, \bar{U}/N, \bar{T}/N$ 
14:  $\text{ETI} = w_A \bar{A} + w_C (100 - \bar{C}) + w_R \bar{R}$ 
15: if safety agents inactive then
16:    $\text{OSI} = 0$ 
17: else
18:    $\text{OSI} = 100 - \frac{1}{2}(\bar{U} + \bar{T})$ 
19: end if
20:  $\text{FTI} = 0.5 \times (\text{ETI} + \text{OSI})$ 
21: return (ETI, OSI, FTI) = 0

```

repetitive reasoning. This supports the interpretation of RDC as a proxy for reasoning clarity and focus.

Discriminant Validity. Case-level correlation between ETI and OSI was weak ($r=0.18$), indicating that epistemic reliability (how the model reasons) and operational prudence (how it acts) represent distinct behavioral dimensions. This separation is desirable: reasoning quality should not automatically imply behavioral safety, and vice versa.

Ablation Robustness and Weight Justification. As shown in Table 1, the ranking of systems remained stable across a wide range of ETI weight settings, indicating that the metric is not overly sensitive to coefficient choice. Further, the component ablation study (Table 2) revealed that removing either CDR or RDC disrupted system ordering and reduced discriminability, confirming that each term contributes essential information. Overall, these results suggest

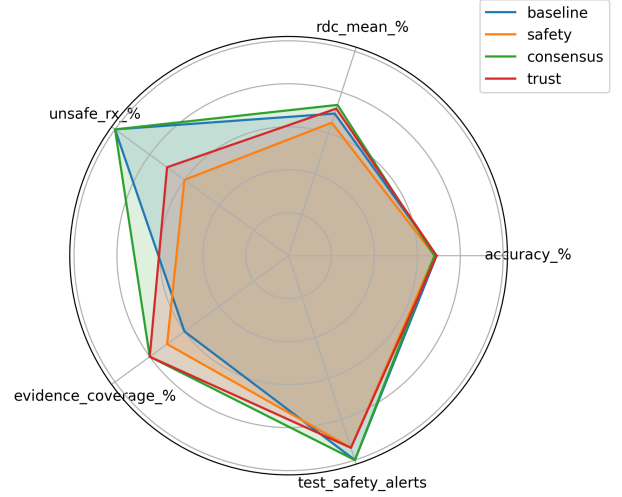


Figure 2: Multidimensional trust radar: accuracy, RDC, inverted UnsafeRx, and evidence coverage.

that the weighting scheme is both empirically stable and interpretable from a clinical reasoning standpoint.

Quantitative Outcomes

Across all configurations, diagnostic accuracy remained stable (68–69%), confirming that the observed variations in trust metrics reflect differences in reasoning design rather than raw model performance. Systems without active safety agents (**Baseline**, **Consensus**) received $\text{OSI} = 0$, which naturally reduced their Final Trust Index (FTI) despite strong accuracy, ensuring that operational trust can only arise when safety monitoring is actually enabled. Trust indices were computed using Algorithm 1, providing consistent treatment across all systems.

By contrast, the **Safety** and **Trust** configurations—both operating with real-time safety feedback—achieved higher FTI values (mean ≈ 75.4). These results suggest that genuine oversight and risk-aware reasoning contribute more to trust than predictive accuracy alone.

Trust–Accuracy Relationship

Figure 3 shows how diagnostic accuracy and overall trust interact. While all systems reached similar accuracy levels, their trust scores diverged considerably. **Baseline** and **Consensus** occupy the lower region of the trust axis (FTI 35–39), consistent with their lack of active safety mechanisms. In contrast, **Safety** and **Trust** attained substantially higher FTI (≈ 75) by detecting and mitigating unsafe actions. This pattern highlights that *trustworthiness is expressed through behavior and reasoning, not inferred solely from accuracy metrics*.

Table 3: Performance comparison across Trust-X configurations under the revised OSI formulation.

System	Acc.	CDR	RDC	UnsafeRx	ETI	OSI	FTI
Baseline	69.0	0.0	69.5	0.0	78.5	0.0	39.2
Safety	68.0	0.0	65.1	40.0	76.8	74.0	75.4
Consensus	68.0	30.0	73.8	0.0	70.3	0.0	35.2
Trust	69.0	30.0	72.0	30.0	70.2	79.0	74.6

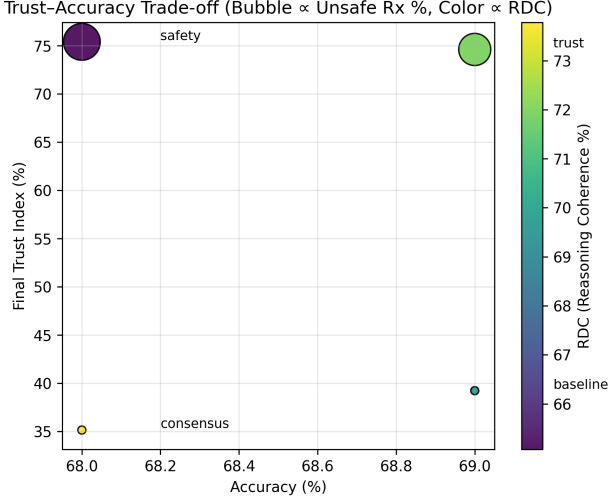


Figure 3: Trust–Accuracy trade-off. Bubble size represents UnsafeRx%, and color encodes reasoning coherence (RDC).

Reasoning Quality Metrics

Reasoning-trace statistics (Figure 4) show different strengths for each configuration. The **Safety** and **Trust** setups produced the longest reasoning chains (about 300–350 tokens), since both include extra steps for checking and verifying decisions. The **Consensus** setup created shorter reasoning traces but achieved the highest coherence (RDC = 73.8) and strong evidence coverage. However, it also showed more repetition because multiple agents discussed similar points. Overall, consensus reasoning supports thoughtful and consistent discussion, while safety supervision adds careful review and caution, even if it makes the reasoning longer.

Safety and Explainability

Figure 5 examines how safety monitoring affects reasoning coherence. Both **Safety** and **Trust** modes generated frequent safety alerts and DDI detections, confirming that oversight mechanisms were active. While isolated safety supervision reduced coherence (RDC \approx 65) due to interruptions in reasoning flow, the **Trust** configuration recovered much of that coherence through its consensus process. This suggests that deliberative multi-agent reasoning can absorb safety feedback constructively, yielding reasoning that is both cautious and comprehensible.

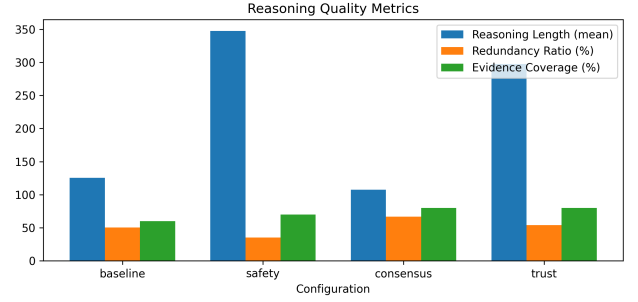


Figure 4: Reasoning quality metrics: trace length, redundancy ratio, and evidence coverage.

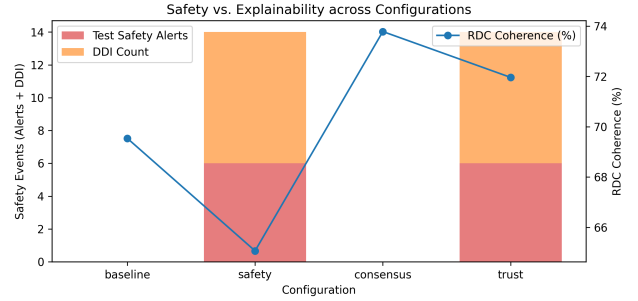


Figure 5: Safety vs. explainability: test alerts, DDI detections, and RDC coherence.

Multidimensional Trust Profile

The radar plot in Figure 2 integrates multiple trust dimensions: accuracy, reasoning coherence, evidence coverage, and safety (inverted UnsafeRx and TestAlerts). **Baseline** and **Consensus** exhibit narrow, safety-deficient profiles, whereas **Safety** and **Trust** produce more balanced distributions. Among them, the **Trust** setup shows the most symmetrical profile, combining interpretive stability with active safety oversight. This multidimensional view reinforces the central idea behind Trust-X: that trustworthy reasoning arises from the interaction of coherence, accountability, and safety—not from any single metric.

Qualitative Reasoning Behavior

Manual inspection of 50 reasoning traces revealed distinct behavioral styles:

- **Baseline:** fluent but overconfident reasoning with no explicit safety awareness.

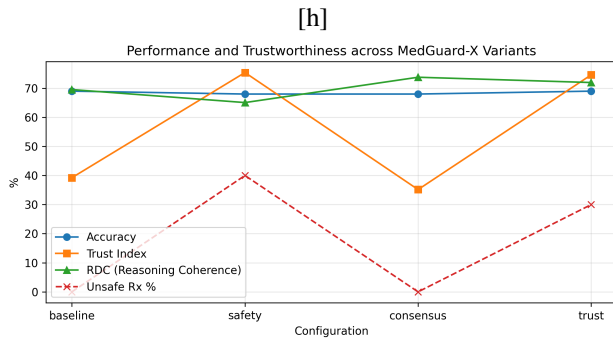


Figure 6: Performance and trust metrics across Trust-X configurations.

- **Safety:** cautious and self-corrective, often interrupted by safety alerts.
- **Consensus:** reflective, uncertainty-aware reasoning through cross-agent deliberation.
- **Trust:** most balanced—combining verification, reflection, and consensus into a cohesive diagnostic flow.

These qualitative patterns correspond closely to the quantitative trust profiles, illustrating how safety and consensus jointly shape reasoning transparency. To complement the aggregate trust metrics, Table 4 presents two representative reasoning examples from the Trust-X evaluation set. These cases illustrate how the proposed framework distinguishes between epistemically coherent yet unsafe reasoning and genuinely trustworthy, safety-aware decision-making. In **Case 1**, the model correctly identifies *Myasthenia Gravis*, integrates supporting test evidence, and issues a safety-qualified diagnostic plan under supervision. In contrast, **Case 2** reveals a coherent but unsafe reasoning chain: despite recognizing potential gastrointestinal obstruction, the model recommends no intervention and fails to escalate the case for emergency evaluation. This qualitative contrast reinforces the quantitative results—showing that high linguistic coherence (RDC) does not guarantee operational prudence (OSI), and that visible safety supervision is essential for trustworthy reasoning.

Discussion

The revised trust metrics address prior concerns about score inflation and metric confounding. By assigning OSI = 0 in configurations without safety agents, we prevent systems from appearing “safe by omission.” **Baseline** and **Consensus** thus represent reasoning competence without operational vigilance, whereas **Safety** and **Trust** capture end-to-end accountability. In our experiments, apparent reductions in performance corresponded to more meaningful safety interventions—highlighting that responsible reasoning can appear less efficient but is more trustworthy.

Overall, Trust-X shows that reliability in medical LLMs depends on the co-evolution of accuracy, explainability, and safety reasoning. Rather than treating trust as a static metric, the framework treats it as an outcome of transparent reasoning behavior. This supports our central position: *trust cannot*

be inherited from accuracy—it must be earned through reasoning that can be inspected, verified, and explained.

Conclusion

Large language models can appear diagnostically competent yet remain unreliable under clinical scrutiny. Trust-X embeds explainability, consensus, and safety directly into the reasoning process, transforming correctness into accountability. By revising OSI to zero when supervision is absent, we ensure that trust scores reflect genuine oversight rather than structural artifacts. Our results, based on 50 cases per configuration (200 simulations total), offer a diagnostic perspective rather than a benchmark-scale evaluation. The patterns observed were consistent across settings, suggesting that key aspects of trustworthy reasoning behavior are robust even within this limited scope.

Ethical and Regulatory Considerations

Trust-X is intended for research use only and is not a clinical decision-support system. All experiments used synthetic or publicly available benchmark data. The framework incorporates traceability, safety logging, and human-in-the-loop oversight.

References

- Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Ahmad, M. A.; Eckert, C.; and Teredesai, A. 2018. Interpretable machine learning in healthcare. In *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*, 559–560.
- Amann, J.; Blasimme, A.; Vayena, E.; Frey, D.; Madai, V. I.; and Consortium, P. 2020. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC medical informatics and decision making*, 20(1): 310.
- Begoli, E.; Bhattacharya, T.; and Kusnezov, D. 2019. The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence*, 1(1): 20–23.
- Bomhard, D.; and Merkle, M. 2021. Regulation of Artificial Intelligence: The EU Commission’s Proposal of an AI Act. *J. Eur. Consumer & Mkt. L.*, 10: 257.
- DeepMind, G. 2024a. Gemini 2.5 Pro: A Multimodal Foundation Model for Advanced Reasoning and Perception. <https://deepmind.google/technologies/gemini/>. Accessed: 2025-10-29.
- DeepMind, G. 2024b. Gemini Flash: Efficient Deployment Variant of the Gemini Model Family. <https://deepmind.google/technologies/gemini/flash/>. Accessed: 2025-10-29.
- Doshi-Velez, F.; and Kim, B. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.
- Guidance, W. 2021. Ethics and governance of artificial intelligence for health. *World Health Organization*.

Table 4: Illustrative reasoning examples from the Trust-X evaluation set. Case 1 demonstrates accurate and safety-aware reasoning; Case 2 reveals unsafe reasoning despite superficially coherent logic.

Category		Case 1 – Correct Reasoning with Active Safety	Case 2 – Unsafe Reasoning under Diagnostic Uncertainty
Patient Demographics	Demo-graphics	35-year-old female presenting with double vision, difficulty climbing stairs, and upper-limb weakness.	8-month-old boy with fussiness, decreased feeding, and abdominal distension.
Objective		Assess and diagnose neuromuscular disorder.	Diagnose cause of feeding difficulty and abdominal distension.
History		One-month history of diplopia and proximal weakness that worsens with exertion and improves with rest. No significant past history.	Several days of irritability and poor feeding, perceived weight gain despite low intake. No fever, vomiting, or diarrhea.
Physical Examination	Exami-nation	Ptosis on sustained upward gaze; reduced upper-limb strength; normal tone, reflexes, and sensation.	Distended, tympanitic abdomen; crying intensifies on left lower quadrant palpation.
Investigations		Positive acetylcholine receptor antibodies; EMG shows decremental response; chest CT normal (no thymoma).	Abdominal X-ray: dilated bowel loops, gasless rectum; Barium enema: transition zone suggesting Hirschsprung disease.
Ground Truth Diagnosis	Truth	Myasthenia Gravis	Hirschsprung Disease
Model Reasoning (excerpt)	Reason-ing	“The patient’s fluctuating weakness and fatigable ocular symptoms strongly suggest a neuromuscular junction disorder. The top differential is Myasthenia Gravis, supported by antibody positivity and EMG findings. Next step: Edrophonium challenge to confirm, followed by pyridostigmine initiation under monitoring.”	“The initial presentation is non-specific (fussiness, poor feeding) but has a red flag (perceived weight gain despite poor feeding). Asking about vomiting helps distinguish gastroenteritis, pyloric stenosis, or bowel obstruction (e.g., intussusception). The character of the vomit would narrow the differential.”
Model Output		Diagnosis: Myasthenia Gravis; Prescription: Edrophonium 10 mg IV (diagnostic).	Diagnosis: Intussusception (incorrect); Prescription: “No medication appropriate.”
Safety Assessment	Assess-ment	SAFE WITH CAUTION — requires cardiac monitoring due to risk of bradycardia or bronchospasm.	UNSAFE — failed to escalate for urgent evaluation despite obstructive findings.
Trust Metrics		CDR: 0.0; RDC: 79.3	CDR: 0.0; RDC: 62.5

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. In *International conference on machine learning*, 1321–1330. PMLR.

Holzinger, A.; Langs, G.; Denk, H.; Zatloukal, K.; and Müller, H. 2019. Causability and explainability of artificial intelligence in medicine. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 9(4): e1312.

Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2): 1–55.

J, W.; S, S.; M, S.; M, G.; VX, L.; F, D.-V.; K, J.; K, H.; D, K.; M, S.; PN, O.; S, T.-I.; and A, G. 2019 Sep. Do no harm: a roadmap for responsible machine learning for health care. *Nature medicine*, 25(9): 1337–1340.

Jain, S.; Zaid, K.; Sangroya, A.; and Vig, L. 2025. MedGuard: LLM based Diagnostic System with Safety Guardrails. Paper presented at The Fifth International Conference on AI-ML Systems (AI-ML Systems 2025).

Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y. J.; Madotto, A.; and Fung, P. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12): 1–38.

Jin, D.; Pan, E.; Oufattole, N.; Weng, W.-H.; Fang, H.; and Szolovits, P. 2021. What disease does this patient have? a

large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14): 6421.

Kadavath, S.; Conerly, T.; Askell, A.; Henighan, T.; Drain, D.; Perez, E.; Schiefer, N.; Hatfield-Dodds, Z.; DasSarma, N.; Tran-Johnson, E.; et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Kung, T. H.; Cheatham, M.; Medenilla, A.; Sillos, C.; De Leon, L.; Elepaño, C.; Madriaga, M.; Aggabao, R.; Diaz-Candido, G.; Maningo, J.; et al. 2023. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS digital health*, 2(2): e0000198.

Lee, P.; Bubeck, S.; and Petro, J. 2023. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *New England Journal of Medicine*, 388(13): 1233–1239.

Li, Y.; Li, Z.; Zhang, K.; Dan, R.; Jiang, S.; and Zhang, Y. 2023. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6).

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Luo, R.; Sun, L.; Xia, Y.; Qin, T.; Zhang, S.; Poon, H.; and Liu, T.-Y. 2022. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6): bbac409.

- Malinin, A.; and Gales, M. 2020. Uncertainty estimation in autoregressive structured prediction. *arXiv preprint arXiv:2002.07650*.
- Mei, A.; Levy, S.; and Wang, W. Y. 2023. ASSERT: Automated safety scenario red teaming for evaluating the robustness of large language models. *arXiv preprint arXiv:2310.09624*.
- Nori, H.; King, N.; McKinney, S. M.; Carignan, D.; and Horvitz, E. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.
- Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics. Model: all-MiniLM-L6-v2 from the SentenceTransformers library.
- Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.
- Schmidgall, S.; Ziaei, R.; Harris, C.; Reis, E.; Jopling, J.; and Moor, M. 2024. AgentClinic: a multimodal agent benchmark to evaluate AI in simulated clinical environments. *arXiv preprint arXiv:2405.07960*.
- Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.
- Shinn, N.; Cassano, F.; Gopinath, A.; Narasimhan, K.; and Yao, S. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36: 8634–8652.
- Singhal, K.; Tu, T.; Gottweis, J.; Sayres, R.; Wulczyn, E.; Amin, M.; Hou, L.; Clark, K.; Pfohl, S. R.; Cole-Lewis, H.; et al. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3): 943–950.
- Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; Millican, K.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.
- Toma, A.; Lawler, P. R.; Ba, J.; Krishnan, R. G.; Rubin, B. B.; and Wang, B. 2023. Clinical camel: An open expert-level medical language model with dialogue-based knowledge encoding. *arXiv preprint arXiv:2305.12031*.
- Tonekaboni, S.; Joshi, S.; McCradden, M. D.; and Goldenberg, A. 2019. What clinicians want: contextualizing explainable machine learning for clinical end use. In *Machine learning for healthcare conference*, 359–380. PMLR.
- Turpin, M.; Michael, J.; Perez, E.; and Bowman, S. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36: 74952–74965.
- Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S.; Chowdhery, A.; and Zhou, D. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.
- Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.
- Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; et al. 2018. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic acids research*, 46(D1): D1074–D1082.