WHEN AGENTS SAY ONE THING AND DO ANOTHER: VALIDATING ELICITED BELIEFS FROM LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) are increasingly deployed in high-stakes settings where good decisions require forming beliefs over the probability of unknown outcomes. However, it is unclear whether LLMs act as if they hold coherent beliefs when making decisions, or if so, how we could validate models' reports of such beliefs. We propose a decision-theoretic framework that elicits both probability judgments and decisions from an agent and tests their mutual consistency. Formally, our methods characterize whether it is possible for the actions to be produced by a "near-rational" decision maker who holds the elicited probability as their true belief. We show that, perhaps surprisingly, this formalization implies empirically testable conditions even without any assumption about the agent's utility function. Applying our framework to stylized clinical diagnosis tasks, we find that models' reported beliefs are demonstrably imperfect summaries of the information revealed in their decisions, but that the discrepancies are small for the strongest models.

1 INTRODUCTION

LLMs are increasingly asked to make decisions under uncertainty, where success requires forming probabilistic beliefs about unknowns in the world. For example, agents for medical tasks must reason about a patient's latent health state in order to trade off the costs and benefits of different actions. How can we tell whether an LLM acts as if it holds coherent beliefs when making decisions? We could prompt the model to output a belief – e.g. the probability that the patient has a particular diagnosis – but it is entirely possible that the model's response is simply an individual text output disconnected from any other output or decision the model might make. Even defining what it could mean for an LLM to hold a belief is highly nontrivial. Nevertheless, validated methods for belief elicitation would be highly useful. For example, elicited beliefs would allow LLMs to provide a legible explanation of why they make particular decisions, an important component of safety and transparency in high-stakes settings. They would also enable diagnosis of failures. For example, a model developer could tell whether a medical agent fails to ask a patient followup questions Liu et al. (2024); Johri et al. (2025) because it holds overconfident beliefs about the diagnosis, or because it perceives a cost to burdening the user with questions, and use this understanding to design appropriate remedies.

Recent work has largely focused on the accuracy and calibration of expressed probabilities (Ulmer et al., 2024; Wang et al., 2024; Cruz et al., 2024). These questions are important, but they are conceptually distinct from the one we study here. On the one hand, a decision maker may hold inaccurate or miscalibrated beliefs in good faith. On the other hand, an LLM may express probabilities that fail to reflect the quantities actually governing its decisions. For example, finetuning or preference optimization may induce the model to verbalize systematically distorted probabilities.

We develop formally-grounded strategies for validating whether probabilistic beliefs elicited from LLMs are a useful explanation for their decisions. Our strategy centers on jointly eliciting beliefs and decisions in tasks where optimal actions depend on the relevant probabilities. We propose two characteristics that elicited beliefs should satisfy in order to be a meaningful description of how an agent makes decisions. First, they should be a *sufficient statistic* which encapsulates all decision-relevant information the agent has about the unknowns. Second, they should be *action-guiding* in the sense that the agent's probability of selecting different actions shifts systematically in response to the elicited belief. We provide a mathematical formalization of these conditions and show how each can

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

be empirically tested in a black-box manner using only data on elicited beliefs and decisions from an agent. We then show a specific sense in which these conditions fully characterize what is possible for black-box testing of elicited beliefs: they are provably necessary and sufficient for there to exist an agent within an expansive class of "near-rational" decision-makers who holds the elicited belief as a true subjective probability and is observationally equivalent to the LLM. Finally, we instantiate this framework empirically across multiple LLM families on clinically grounded diagnostic tasks spanning four medical settings. We find that elicited beliefs are provably imperfect, in that models have additional decision-relevant information about a patient's diagnosis which is revealed in their actions but not communicated in the elicited belief. However, the degree of discrepancy is relatively small for the strongest models, indicating that they are – at least observationally – close to an agent whose actions are coherently attributable to their elicited beliefs.

2 Related Work

Belief elicitation from humans: A long literature in psychology and economics seeks to elicit subjective probabilities, often using monetary incentives to encourage truthful reporting. Reviews Schlag et al. (2015); Charness et al. (2021) emphasize that validation is difficult and arguably ill-posed: beliefs are unobservable and may be shaped by elicitation itself, so methods are often judged by how well elicited beliefs predict actions. Our approach is similar in spirit, but LLMs pose different constraints. Human studies typically assume preferences, such as more money being better, to design proper scoring rules Savage (1971); Gneiting & Raftery (2007) or test best responses in games Nyarko & Schotter (2002); Rey-Biel (2009). For LLMs, assuming a specific utility is strong and hard to justify, and credibly incentivizing a model in simulation is difficult. We therefore remain agnostic about the model's utility. Finally, whereas belief elicitation for humans can "contaminate" subsequent actions Croson (2000); Blanco et al. (2010), we avoid this for LLMs by eliciting beliefs and actions in separate context windows.

Belief elicitation from LLMs: With the advent of large language models there has been renewed interest in characterizing LLM beliefs. Herrmann & Levinstein (2024) provide a philosophical account, and our work can be seen as operationalizing their "use" criterion. Other work has studied how to elicit priors from LLMs Zhu & Griffiths (2024) and whether LLMs express beliefs that are consistent with probabilistic axioms Zhu & Griffiths (2025); Freedman & Toni (2025). Our work introduces a new set of tools to test coherence of beliefs in comparison to decisions. Perhaps the most related is work by Pal et al. (2025), who test whether LLMs make bets on events which go in the same direction as their beliefs. However, they do not introduce a formal framework for testing when beliefs and actions should correspond in a specific way. Further afield, an emerging literature tests whether LLMs' outputs are consistent with various cognitive biases seen in humans Echterhoff et al. (2024); Binz & Schulz (2023); Cheung et al. (2025) but this work does not focus on validity of elicited beliefs.

Mechanistic interpretability: A central motivation for our study is the concern that LLMs may know more than what they reveal in verbalized probabilities. Mechanistic interpretability uses access to models' internal representations to test such questions Sharkey et al. (2025); Bereska & Gavves; for example, probes can predict distinctions such as true vs. false claims from activations Azaria & Mitchell (2023); Marks & Tegmark (2023). We view this work as complementary: our tests are fully black-box and output-only, making them useful for closed models, while many mechanistic approaches (e.g., linear probes) require supervised labels, injecting external information and complicating a direct interpretation of their outputs as beliefs of the model. Given the difficulty of formally defining "belief," both toolkits are valuable.

Medical applications: A growing body of work evaluates large language models in clinically consequential settings, emphasizing that downstream quality depends on decision policies rather than static accuracy alone (Singhal et al., 2025; Gaber et al., 2025; Hager et al., 2024; Williams et al., 2025). Clinical evaluations further stress that models may lack reliable metacognitive awareness of their own limitations, making naive self-reported confidence insufficient for safe triage (Griot et al., 2025; Hager et al., 2024). Evaluations have also stressed difficulties in seeking information when necessary to make a diagnosis, perhaps reflecting incorrect probabilistic reasoning Johri et al. (2025); Li et al. (2024). Our aim is to provide a principled framework within which to understand how models' understanding of uncertainty relates to their decisions.

3 METHODOLOGY

3.1 DECISION-THEORETIC FRAMEWORK

We formalize the decision problem as follows. An environment generates an unknown state of the world $\theta \sim P^\star(\theta)$ and observations $x \sim P^\star(x \mid \theta)$. The induced *ground-truth* posterior under the environment is $P^\star(\theta \mid x)$. After observing x, an agent forms its own (potentially misspecified) *subjective* posterior belief $P_S(\theta \mid x)$, which need not equal $P^\star(\theta \mid x)$. In order for a subjective probability to have empirically testable content, it must be related to the agent's actions in some way. In the classic framework of expected utility maximization, the agent selects an action $a \in \mathcal{A}$ to maximize expected utility: $a(x) = \arg\max_{a \in \mathcal{A}} \sum_{\theta \in \Theta} P_S(\theta|x)u(a,\theta)$. This optimization requires two components: (1) *beliefs* about the unknown state—captured here by the subjective posterior $P_S(\theta \mid x)$—and (2) *preferences* over outcomes—captured here by the utility u. In this paper, our goal is to formulate empirically testable conditions for an agent's *reported* belief to represent a subjective probability in the sense that it represents a real account of why the agent acts the way that it does. Expected utility maximization is one framework linking beliefs to actions, but may represent too stringent a formal model on its own – an agent could be usefully described as having beliefs even if it fails to exactly maximize a specific utility function. We instead propose conditions which can be justified in a broader sense. While we argue that these conditions are independently meaningful apart from any specific mathematical model, they can be formally grounded in an expansive class that captures "near-rational" decision making. Let $q_a(x)$ denote the probability that the agent takes action a after observing x.

Definition 1. *An agent is a perturbed utility maximizer if, for some utility function $u : \mathcal{A} \times \Theta \to \mathbb{R}$ that is non-degenerate in relative action utilities (i.e., there exist x, x' in the support of X and distinct $a, a' \in \mathcal{A}$ such that $\sum_{\theta \in \Theta} P_S(\theta \mid x)[u(a,\theta) - u(a',\theta)] \neq \sum_{\theta \in \Theta} P_S(\theta \mid x')[u(a,\theta) - u(a',\theta)]$) and some strictly convex regularizer $C : \Delta(\mathcal{A}) \to \mathbb{R}$, its choice-probability vector $q(x) \in \Delta(\mathcal{A})$ satisfies*

$$q(x) \in \arg\max_{\lambda \in \Delta(\mathcal{A})} \left\{ \sum_{a \in \mathcal{A}} \lambda_a \left(\sum_{\theta \in \Theta} P_S(\theta \mid x) u(a,\theta) \right) - C(\lambda) \right\}.$$

The realized action is then drawn according to this choice rule: $\Pr(A = a_i \mid x, \theta) = q_{a_i}(x)$.

Perturbed utility maximization originated in the economics literature as a generalization of traditional choice models (Hofbauer & Sandholm, 2002). The regularizer C allows this class to express a wide variety of near-utility maximizing behaviors where agents act with stochasticity, with unobserved components in their utility function, or to ensure robustness to different preferences. For example:

Example 1. *If $C(q) = \sum_{a \in \mathcal{A}} q_a(x) \log q_a(x)$ is the negative entropy, then the choice probabilities are a softmax over expected utilities, i.e. the well-known logit model (McFadden, 1973; Train, 2009).*

Example 2. *More generally, Hofbauer & Sandholm (2002) show any random utility model satisfies Definition 1. Random utility models are the most common class of discrete choice models, and model the agent as maximizing $u(a,\theta) + \epsilon_a$ for a random shock ϵ_a to account for unobserved variables impacting the agent's preferences.*

Example 3. *Taking q_0 to be a baseline policy and $C = D_{KL}(q||q_0)$ to be the KL divergence from q_0 recovers an agent who trades off expected utility against deviation from a default policy (as in many post-training pipelines for LLMs (Ouyang et al., 2022)).*

Other examples occur across the economic literature; e.g., Fudenberg et al. (2015) show that Definition 1 encompasses ambiguity-averse agents that maximize their worst-case expected utility over a set of different scenarios for the utility function, while Fosgerau et al. (2020) show that it can be used to express common models of rational inattention. Accordingly, formally grounding empirical tests of belief consistency with respect to Definition 1 allows us to express properties that should relate beliefs and actions for a very broad set of agents that deliberately pursue goals in some coherent way, without requiring strict expected utility maximization. One deliberate choice is that we restrict the utility u to depend only on a and θ, not x. Otherwise any action distribution $a(x)$ is trivially rationalizable (e.g., by taking $u(a,\theta,x) = 1[a = a(x)]$) without requiring the agent to have any beliefs about θ at all. Effectively, there is no observable distinction between an agent whose preferences correlate with θ via x and an agent who has beliefs about θ and so we choose to interpret the component of the agent's decision-making which is able to encode predictive information about θ as part of the belief state.

3.2 TESTABLE IMPLICATIONS FOR BELIEF ELICITATION

A central objective of this paper is to test whether it is possible for a model's verbalized probabilities to reflect *subjective* belief states that meaningfully drive its choices. We will refer to the probability judgment *elicited* from a model by some method (e.g., prompting) as $P_E(\theta \mid x)$. Our goal is to test whether it is possible for a (near) rational decision maker to hold those judgments as true subjective beliefs, $P_S(\theta|x) = P_E(\theta \mid x)$, and still take the same actions as the model. We will construct a framework to test this in the context of a specific decision problem, where we assume that the experimenter has samples from the true distribution over (x, θ).

We gather data in two steps. First, on a sample of values of x, we elicit $P_E(\theta \mid x)$ using whatever method we wish to test. In our experiments, we prompt the model to report a probability distribution $P_E(\theta \mid x)$ in natural language. Specifically, for each context x in our evaluation set, we query the model for probability estimates for each possible state. For the tasks we consider, the state space is binary ($\theta \in \{0, 1\}$, indicating absence or presence of a condition), so we elicit $p(x) := P_E(\theta = 1 \mid x)$ directly as a numerical probability. However, our methods apply to any elicitation method. We focus on natural-language probability estimates because of their simplicity and the wide literature investigating them. For notation, we often use $p(x) := P_E(\theta = 1 \mid x)$ or $p^*(x) := P^*(\theta = 1 \mid x)$ in shorthand for the binary case; we use $P_E(\theta \mid x)$ or $P(\theta \mid x)$ when results generalize beyond the binary case. Second, we prompt the model with the same x in a separate context and ask it to take an action a. This generates a sample of the form $(x_i, P_E(\theta = 1 \mid x_i), a_i, \theta_i)$, where we observe for each x_i the model's elicited belief, its action, and the true state (which was not revealed to the model). Finally, we apply tests examining properties that the sample distribution should satisfy if the elicited beliefs play a meaningful role in action. Violations quantify the extent to which elicited beliefs depart from a useful interpretation as true subjective probabilities. We next detail these tests.

Conditional independence of actions and outcomes: Intuitively, a valid report of a belief should summarize all of the decision-relevant information that the agent possesses about the state θ. Formally, once we condition on the model's subjective belief, the agent's action a should not provide any additional information about the true outcome θ: P_S is a sufficient statistic for the impact of the distribution of future events on the agent's choices. If a rational agent's actions do have excess correlation with the true state given their elicited P_E, the agent must "know more than they say".

Proposition 1 (Conditional Independence under Truthful Reporting)**.** *For any agent satisfying Definition 1, their subjective beliefs and actions satisfy $a \perp \theta \mid P_S(\theta \mid x)$. Conversely, if for a set of elicited beliefs P_E, $a \not\perp \theta \mid P_E(\theta|x)$, then there is no agent satisfying Definition 1 (under* any *utility function u and regularizer C) who holds subjective belief $P_S = P_E$.*

For intuition, note that in Definition 1, the expected utility term depends on x only through $P_S(\theta|x)$. Once this belief term is fixed, sampling $a \sim q(x)$ conveys no additional information about θ.

Remark 1. Observing $a \perp\!\!\!\perp \theta \mid P_E(\theta \mid x)$ does *not* by itself suffice to prove truthful reporting by a rational agent. E.g. consider an agent who takes a uniformly random action independent of x and reports a (independent) uniformly random value in [0,1] as $p_E(x)$. This satisfies $a \perp \theta \mid P_E(\theta|x)$ without satisfying Definition 1.

Remark 2. While conditional independence can be derived from Definition 1, it can be shown to hold even more broadly. For example, we show in the appendix that it holds for agents described by prospect theoretic risk-aversion (Kahneman & Tversky, 1979).

In order to empirically test this condition, we first conduct a direct conditional-independence (CI) test of the null hypothesis $A \perp\!\!\!\perp \theta \mid P_E(\theta \mid x)$ via the nonparametric CMI (Conditional Mutual Information) test for $H_0 : I(A; \theta \mid p) = 0$. While a binary reject/fail-to-reject CI decision is useful, it can be hard to interpret *how large* a violation is in practice. Therefore, we additionally quantify the *degree* of CI violation via an out-of-sample predictive-performance comparison. Concretely, we train two predictive models for the realized state θ and compare nested feature sets. *Model 1* conditions on the elicited belief $p_E(x)$, while *Model 2* conditions on $(A, p_E(x))$. Under the conditional-independence null, adding the action A should not improve prediction of θ once we condition on p_E. We use cross-validated mean squared error (MSE) and report percent improvement as the relative reduction in MSE from including A. Significance is assessed by a bootstrap CI.

Testing whether beliefs move actions. For beliefs to meaningfully describe behavior, they should be *action-guiding*: actions should shift systematically as beliefs change. For an expected utility maximizer, expected utilities are linear in probabilities, yielding action-switching thresholds. This deterministic structure is too strong when choices are stochastic or perturbed, so we instead require choice probabilities to move monotonically with beliefs. With multiple actions, this cannot generally be imposed action-by-action, since actions compete for probability mass. We therefore use the convex-analytic monotonicity condition implied by perturbed utility maximization in Definition 1. Let $b \in \Delta(\Theta)$ be a belief over states, $q(b) := (q_1(b), \ldots, q_J(b))$ the conditional choice-probability vector, with $q_a(b) := \Pr(a(x) = a \mid P_S(\cdot \mid x) = b)$, and $v(b) \in \mathbb{R}^J$ the expected-utility index, with $v_a(b) := \sum_{\theta \in \Theta} b_\theta u(a, \theta)$.

Proposition 2 (Cyclic monotonicity)**.** *For any agent satisfying Definition 1, the conditional choice-probability map $q(b)$ is cyclically monotone with respect to $v(b)$. That is, for any finite collection $b_1, \ldots, b_M$ with $b_{M+1} = b_1$, $\sum_{m=1}^{M} q(b_m) \cdot \big(v(b_m) - v(b_{m+1})\big) \geq 0$. Consequently, if the observed conditional choice probabilities violate these inequalities, then the choices are not rationalizable by any agent in the model class.*

Cyclic monotonicity generalizes standard monotonicity to settings with multiple actions and choice-probability vectors. It was previously known to hold for random utility models (McFadden & Richter, 1990) and is easily shown to hold for the more general perturbed maximization class via convex-analytic arguments (Rockafellar, 1997). In general, the condition can be somewhat abstract, but we show it has a simpler interpretation when $\theta \in \{0, 1\}$. Define $d_a := u(a, 1) - u(a, 0)$ and let $q_a(p) := \Pr(a(x) = a \mid p(x) = p)$, where $p(x) = P_E(\theta = 1 \mid x)$.

Proposition 3 (Binary-state characterization of cyclic monotonicity)**.** *Suppose $\theta \in \{0, 1\}$. For the utility-difference vector d induced by u, the conditional choice rule $q(p)$ is cyclically monotone with respect to the utility index $v(p)$ if and only if, for every $p < p'$,*

$$\sum_{a=1}^{J} q_a(p) d_a \leq \sum_{a=1}^{J} q_a(p') d_a.$$

Moreover, under Definition 1, the inequality is strict whenever $q(p) \neq q(p')$.

Thus, as the elicited probability of the positive state increases, the model's choice probabilities should move toward actions that are relatively more valuable in state 1. To test this implication, we divide elicited beliefs into K quantile bins, with bin centers $\bar{p}_1 \leq \cdots \leq \bar{p}_K$. For each bin k, we estimate the vector of action shares $\widehat{q}_k = (\widehat{q}_{1k}, \ldots, \widehat{q}_{Jk})$. Define $\widehat{m}_k(d) := \sum_{a=1}^{J} \widehat{q}_{ak} d_a$, where $d_a := u(a, 1) - u(a, 0)$ is the relative value of action a in state 1 rather than state 0. By Proposition 3, the data are consistent with the strict implication of the model if there exists a utility-difference vector $d \in \mathbb{R}^J$ such that $\widehat{m}_k(d)$ is weakly increasing in k, and is strictly increasing across any adjacent bins $k, k+1$ for which $\widehat{q}_k \neq \widehat{q}_{k+1}$. Only the relative values of d matter because each $\widehat{q}_k$ is a probability vector. We therefore normalize d by setting one coordinate equal to 0 and another equal to 1. Since we do not know which actions have the smallest and largest state-1 relative utilities, we try all ordered pairs of distinct actions $(a^\star, b^\star)$. Let $\mathcal{S} := \{k \in \{1, \ldots, K-1\} : \widehat{q}_k \neq \widehat{q}_{k+1}\}$ denote the adjacent bin pairs for which the estimated choice-share vectors differ. For each ordered pair $(a^\star, b^\star)$, we solve the signed-margin linear program

$$\begin{aligned} \mathrm{SignedLP}(a^\star, b^\star): \quad & \max_{d, \gamma} \quad \gamma \\ & \text{s.t.} \quad \widehat{m}_{k+1}(d) - \widehat{m}_k(d) \geq \gamma, \quad k \in \mathcal{S}, \\ & \qquad d_{a^\star} = 0, \qquad d_{b^\star} = 1, \\ & \qquad d \in [0, 1]^J, \\ & \qquad \gamma \in \mathbb{R}. \end{aligned} \tag{1}$$

We summarize the signed monotonicity statistic by $\widehat{\gamma} := \max_{a^\star \neq b^\star} \mathrm{val}\big(\mathrm{SignedLP}(a^\star, b^\star)\big)$. This statistic is the largest uniform lower bound on the adjacent-bin increases of the fitted utility-difference index. Hence positive values indicate strict monotonicity with a uniform margin, zero indicates weak but not strict monotonicity, and negative values indicate a violation of weak monotonicity.

Proposition 4 (Equivalence with Proposition 3)**.** *Assume $\mathcal{S} \neq \varnothing$. Then $\widehat{\gamma} > 0$ if and only if there exists a non-constant $d \in \mathbb{R}^J$ such that $\widehat{m}_1(d) \leq \cdots \leq \widehat{m}_K(d)$, and, moreover, $\widehat{m}_k(d) < \widehat{m}_{k+1}(d)$*

for every $k \in \mathcal{S}$. Moreover, $\widehat{\gamma} \geq 0$ if and only if there exists a non-constant $d \in \mathbb{R}^J$ such that $\widehat{m}_1(d) \leq \cdots \leq \widehat{m}_K(d)$. Finally, $\widehat{\gamma} \in [-1, 1]$.

Proposition 4 follows from the normalization used in the LP and the finiteness of $\mathcal{S}$. If a non-constant vector d' satisfies the required weak and strict inequalities, then $\min_{k \in \mathcal{S}}\{\widehat{m}_{k+1}(d') - \widehat{m}_k(d')\} > 0$. The affine rescaling $\widetilde{d}_a = (d'_a - \min_b d'_b)/(\max_b d'_b - \min_b d'_b)$ preserves the inequalities, lies in $[0, 1]^J$, and attains both 0 and 1, so it is feasible for the relevant signed-margin LP with some $\gamma > 0$. The converse is immediate from the LP constraints. The same argument with weak inequalities gives $\widehat{\gamma} \geq 0$ if and only if some normalized non-constant d makes the fitted index weakly increasing. Adjacent pairs outside $\mathcal{S}$ impose no additional restriction, because $\widehat{q}_k = \widehat{q}_{k+1}$ implies $\widehat{m}_{k+1}(d) - \widehat{m}_k(d) = 0$ for every d. Thus $\widehat{\gamma} < 0$ means that every normalized d generates a negative increment for at least one adjacent pair in $\mathcal{S}$. Finally, since $d \in [0, 1]^J$ and each $\widehat{q}_k$ is a probability vector, $\widehat{m}_k(d) \in [0, 1]$, so every adjacent increment lies in $[-1, 1]$ and therefore $\widehat{\gamma} \in [-1, 1]$.

Characterizing the strength of black-box belief elicitation. We now show that the conditional independence and monotonicity conditions constitute the full testable implications of an agent holding a belief, at least in the regime we study with black-box access and no assumptions on the agent's utility function. Formally, while we have already shown that both conditions above are necessary for Definition 1 (perturbed utility maximization) to hold, we now show that they are jointly *sufficient* as well: when both conditions are satisfied, the agent is *observationally equivalent* to some perturbed utility maximizer who holds $P_E(\theta|x)$ as their true subjective probability.

Proposition 5 (Joint characterization of rationalizable elicited beliefs)**.** *Let $\mathbb{P}$ denote the observed distribution over $(P_E(\theta \mid x), \theta, a)$, and let $q(P_E) := \left[\Pr(a = a' \mid P_E(\theta \mid x) = P_E)\right]_{a' \in \mathcal{A}}$ be the induced conditional choice rule. There exists a utility function $u : \mathcal{A} \times \Theta \to \mathbb{R}$ that is non-degenerate in relative action utilities and an agent satisfying the perturbed-maximization representation in Definition 1, except with some convex, but not necessarily strictly convex, regularizer $C : \Delta(\mathcal{A}) \to \mathbb{R}$, whose subjective belief satisfies $P_S(\theta \mid x) = P_E(\theta \mid x)$ and whose induced distribution over $(P_E(\theta \mid x), \theta, a)$ equals $\mathbb{P}$, if and only if $a \perp\!\!\!\perp \theta \mid P_E(\theta \mid x)$ and $q(P_E)$ is cyclically monotone with respect to the utility index $v_a(P_E) = \sum_{\theta \in \Theta} P_E(\theta \mid x)u(a, \theta)$.*

Our theoretical results show a perhaps surprising fact: elicited beliefs have empirically testable implications even after discarding much of the structure underlying traditional belief elicitation for humans. We do not assume a known utility function or require exact optimization. By contrast, the economic and algorithmic literature on proper scoring rules and related elicitation mechanisms typically requires substantive assumptions about incentives (Savage, 1971; Gneiting & Raftery, 2007; Waggoner & Chen, 2014). Our tests instead leverage a distinctive feature of LLMs: they can be repeatedly queried in fresh context windows, allowing direct estimation of choice probabilities. At the same time, our results characterize the limits of this black-box access model. Passing our conditions only shows that, on the evaluated distribution, the agent is observationally indistinguishable from some rational decision maker who truly holds the reported belief. This is closer to failing to falsify the belief report than to proving that the model "truly" holds that belief. For example, the tests would not rule out deliberate misreporting or deception by a sufficiently sophisticated agent.

Consistency across decision tasks. So far, we have discussed testable implications within a single decision task. Across tasks, a rational agent should separate beliefs from preferences: when multiple tasks reference the same state θ, agents satisfying Definition 1 should act as if driven by a common subjective belief p_S, even if utilities or action sets differ. This cross-task stability—often called *prize independence*—is a foundational implication of subjective probability (Anscombe & Aumann, 1963; Savage, 1972; Ronayne et al., 2022). Accordingly, we compare elicited probabilities across prompts that share the same (x, θ) but differ in task framing. In the binary setting, the alternative prompt asks for $p_E(\theta \mid x)$ under an explicitly stated mean-squared-error (MSE) evaluation loss (chosen as MSE is a proper scoring rule that incentivizes telling the true belief). Let $p(x; \pi, j)$ denote the reported probability for context x under prompt π on repetition j, with π_0 the main diagnostic prompt and π_{MSE} the MSE prompt. We compare paired reports within the same repetition and summarize cross-task inconsistency as $\Delta_{\text{MSE}} = \sqrt{\frac{1}{nr}\sum_{i=1}^{n}\sum_{j=1}^{r}\left(p(x_i; \pi_{\text{MSE}}, j) - p(x_i; \pi_0, j)\right)^2}$. Larger values indicate less stable elicited beliefs across task framings. As a reference, we also report within-prompt variability for the standard prompt π_0, computed from repeated elicitations of $p(x; \pi_0, j)$.

Table 1: **Conditional-independence (belief sufficiency) tests.** The first section reports kNN conditional mutual information $I(A;\theta \mid p)$ with 95% confidence intervals. The second reports Random Forest comparisons of $\theta \sim p$ vs. $\theta \sim (A, p)$, summarized as percent MSE improvement with 95% bootstrap CIs. The third reports average RF improvement by model and by dataset/domain.

	kNN CMI		RF improvement		Average RF improvement		
Dataset / Model	**CMI**	**95% CI**	**% Imp.**	**95% CI**	**Group**	**% Imp.**	**95% CI**
Heart–GPT-Min	0.1454	[0.1119, 0.1789]	18.39	[0.99, 31.66]	GPT-Minimal	6.15	[-1.22, 13.52]
Heart–GPT-High	0.0753	[0.0422, 0.1085]	9.93	[2.91, 17.75]	GPT-High	4.89	[1.25, 8.53]
Heart–Llama	0.0718	[0.0365, 0.1070]	14.60	[6.15, 22.64]	Llama	15.19	[6.11, 24.27]
Heart–DeepSeek	0.0675	[0.0354, 0.0997]	22.41	[9.64, 32.98]	DeepSeek-R1	5.11	[-4.83, 15.06]
Cry–GPT-Min	0.2232	[0.1874, 0.2589]	6.27	[-0.40, 11.43]	Heart Disease	16.33	[11.81, 20.85]
Cry–GPT-High	0.1901	[0.1390, 0.2412]	6.92	[2.22, 10.67]	Cry	5.21	[-0.03, 10.45]
Cry–Llama	0.4223	[0.3764, 0.4681]	11.12	[2.39, 20.25]	Fever	2.07	[0.30, 3.85]
Cry–DeepSeek	0.1745	[0.1265, 0.2225]	-3.48	[-6.17, -0.62]	Diabetes	7.73	[-4.93, 20.38]
Fever–GPT-Min	0.1446	[0.1128, 0.1765]	-0.06	[-0.24, 0.14]			
Fever–GPT-High	0.0944	[0.0564, 0.1323]	1.85	[0.32, 3.27]			
Fever–Llama	0.3289	[0.2893, 0.3686]	4.95	[-1.31, 10.65]			
Fever–DeepSeek	0.2060	[0.1663, 0.2456]	1.54	[-3.65, 6.52]			
Diabetes–GPT-Min	0.0193	[0.0129, 0.0258]	0.00	[-0.00, 0.00]			
Diabetes–GPT-High	0.0461	[0.0182, 0.0740]	0.85	[-0.26, 1.91]			
Diabetes–Llama	0.2695	[0.2357, 0.3033]	30.07	[20.56, 38.96]			
Diabetes–DeepSeek	0.0351	[0.0169, 0.0533]	-0.03	[-0.10, 0.01]			

Internal Consistency via Law of Iterated Expectation. So far, we have focused purely on the link between beliefs and actions. However, a parallel literature investigates whether elicited beliefs from LLMs constitute a coherent probability distribution, in the sense of adhering to probability axioms (Zhu & Griffiths, 2025; Herrmann & Levinstein, 2024). To investigate the relationship between these two senses of coherence (decision-theoretic and probabilistic), we develop a test generalizing previous work to the settings we consider. In particular, we test adherence to the law of iterated expectation over an auxiliary variable z (e.g., a patient attribute) that we can optionally reveal. Concretely, let x denote the base context (with z withheld), and let $\{B_1, \ldots, B_k\}$ be a partition of the support of z. We verify: $P_E(\theta \mid x) = \sum_{j=1}^{k} P_E(\theta \mid x, z \in B_j)\ P_E(z \in B_j \mid x)$. In our experiments, we construct partitions based on patient covariates and elicit three quantities (with m standing for model): $P_E(\theta \mid x)$, $P_E(\theta \mid x, z \in B_j)$ for each bin B_j, and $P_E(z \in B_j \mid x)$. We then measure the discrepancy: $\Delta_{\text{LIE}}(x) = \left| P_E(\theta \mid x) - \sum_{j=1}^{k} P_E(\theta \mid x, z \in B_j)\ P_E(z \in B_j \mid x) \right|$. Large values of $\Delta_{\text{LIE}}(x)$ indicate internal inconsistencies in the model's reported beliefs, suggesting they may not reflect a coherent probabilistic reasoning process. We do not expect this sense of coherence to necessarily align with our account of valid belief elicitation. Intuitively, a decision maker can "truthfully" hold beliefs that are not fully consistent in a probabilistic sense (e.g., they could satisfy Definition 1 without P_S being a valid probability distribution). We measure Δ_{LIE} to provide an empirical comparison of where these senses of coherence overlap and differ.

4 Experimental Setup

We apply our methodology to four medical diagnosis tasks: two using real-world datasets (structural heart disease and diabetes) and two using expert-constructed Bayesian networks (pediatric medicine). We evaluate each dataset on the following language models: GPT-5 Thinking High Reasoning and Minimal Reasoning Models (Singh et al., 2025), Deepseek R1 671B (Guo et al., 2025), and Llama-4 Scout 17B (MetaAI, 2024). These models allow us to examine a range of both reasoning levels and model sizes, encompassing both SOTA frontier and open-source models. For each metric, we report 95% bootstrapped CIs. Additional details regarding prompts, datasets and other implementation details can be found in the Appendix and released code. For each dataset, we sample 200 covariate-outcome pairings with five repetitions.

Datasets. We use publicly available datasets: (1) electrocardiograms with demographic data and structural heart disease labels (Elias & Finer, 2025), and (2) diabetic patient records containing fields such as exercise and glucose levels (CDC, 2017). Ground-truth $p^*(x)$ are computed as the fraction of positive diagnoses falling into covariate strata (always at least 100 patients). We also evaluate on distributions from two Bayes nets constructed by a leading expert in pediatric medicine. Pending completion of a data-sharing agreement, we plan to have these permissions by the camera-ready. We

Table 2: Signed-margin monotonicity statistic $\widehat{\gamma}$ by dataset and model. Positive values indicate strict monotonicity with a uniform margin, values equal to zero indicate weak but not strict monotonicity, and negative values indicate violations of weak monotonicity. Brackets report 95% confidence intervals. The right panel reports unweighted averages by model and by domain.

Dataset / Model	$\widehat{\gamma}$ [95% CI]
Heart–GPT-Min	0.104 [0.012, 0.168]
Heart–GPT-High	0.059 [0.000, 0.160]
Heart–Llama	0.042 [0.005, 0.068]
Heart–DeepSeek	0.192 [0.117, 0.205]
Cry–GPT-Min	0.050 [0.001, 0.117]
Cry–GPT-High	0.059 [0.003, 0.120]
Cry–Llama	0.266 [0.217, 0.317]
Cry–DeepSeek	0.097 [0.034, 0.142]

Dataset / Model	$\widehat{\gamma}$ [95% CI]
Fever–GPT-Min	0.053 [0.000, 0.082]
Fever–GPT-High	0.103 [0.001, 0.194]
Fever–Llama	-0.011 [-0.018, -0.005]
Fever–DeepSeek	0.148 [0.078, 0.175]
Diabetes–GPT-Min	0.000 [0.000, 0.032]
Diabetes–GPT-High	0.031 [0.000, 0.049]
Diabetes–Llama	0.000 [0.000, 0.003]
Diabetes–DeepSeek	0.010 [0.000, 0.028]

Average	$\widehat{\gamma}$
GPT-Min	0.052
GPT-High	0.063
Llama	0.074
DeepSeek	0.112
Heart	0.099
Cry	0.118
Fever	0.073
Diabetes	0.010

use networks in pediatric medicine constructed separately for chief complaints of fever and crying, respectively. The fever network contains demographic and symptom covariates (e.g., jaundice). The crying network covers causes such as colic and gas pain with behavioral/physical covariates.

Decision task. For the medical diagnosis tasks that we consider, the state space is binary ($\theta \in \{0, 1\}$ indicating absence or presence of a condition). We prompt the language model with each context x_i and observe its chosen action $a_i \in \mathcal{A}$. For our diagnostic tasks, the action space consists of three options: diagnose the patient as having the condition, diagnose the patient as not having the condition, or 'defer", refusing to make any diagnosis. The model is instructed to first choose if it feels capable of making a diagnostic decision, and then asked which decision the model would make if it had to make a decision with only binary $a = \{1, 0\}$ options presented. These questions can then translate our decision into one of the three options. Deferral is treated as a distinct action with its own cost. When eliciting either beliefs or decisions, we provide the model with the clinical evidence x and specify the relevant outcome θ, then instruct it to return a numerical probability estimate. For example, we might ask for the probability of structural heart disease given that a patient is over 50 and has a particular ECG abnormality, and separately ask the model what diagnostic action it would take from the same evidence. Throughout, we elicit beliefs as $p(x) := P_E(\theta = 1 \mid x)$. Unless otherwise noted, we use a standard version of these prompts without additional instructions.

5 Results

Conditional Independence Test. We first test whether the elicited belief is *decision-sufficient* for the model's actions, as implied by Proposition 1. Under truthful reporting with outcome-dependent loss and exogenous decision noise, the model's action A should add no information about the realized outcome θ once we condition on the elicited belief, equivalently $H_0 : I(A; \theta \mid p) = 0$. The left side of Table 1 reports a kNN ($k = 3$) conditional mutual information (CMI) test; all model–dataset pairs reject the null, with 95% bootstrap CIs (500 resamples) strictly above zero. Appendix Table 4 shows the same conclusion for isotonic-calibrated beliefs, indicating that calibration does not restore decision-sufficiency. To gauge the size of this residual dependence, the right side of Table 1 reports a cross-validated Random Forest comparison, with parameter details in the Appendix to prevent data contamination and control overfitting. We compare a baseline model predicting the realized state from the elicited belief alone, $\theta \sim p$, against an augmented model that also includes the action, $\theta \sim (A, p)$. Under the conditional-independence null, adding A should not reduce out-of-sample MSE once p is included. Empirically, MSE reductions are heterogeneous by both domain and model, with the clearest reductions for the Heart Disease domain and for Llama across domains. These results suggest that p captures most—but not all—decision-relevant signal in many settings; in cases such as Llama and Heart Disease, elicited probabilities do not appear to serve as sufficient statistics for decisions, motivating task- and model-specific validation.

Cyclic Monotonicity Test. Table 2 reports the signed cyclic-monotonicity margin $\widehat{\gamma}$ for each dataset/model pair, computed by binning elicited probabilities into five quantiles and estimating choice shares within bins. Under this formulation, $\widehat{\gamma}$ is the largest uniform lower bound on adjacent-bin increases in the fitted utility-difference index: positive values indicate strict monotonicity, zero indicates weak but not strict monotonicity, and negative values indicate a violation of weak monotonicity. Results vary across domains and models. In 9 of 16 dataset/model combinations,

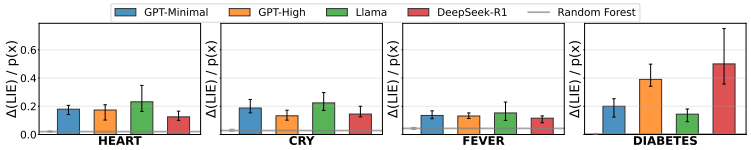

Figure 1: Internal Consistency Error via Law of Iterated Expectation (LIE): Here we plot the 95% CI for the median quantity $\Delta_{\text{LIE}}(x)/p(x)$ (see Section 3.2) by dataset(subplots) and model(bars). A cross-validated random forest model that is separately trained on ground truth $p^*(x)$ and $p^*(x, z)$ (also conditional on next state distribution z) values is used as a baseline.

the 95% confidence interval lies strictly above zero, indicating strict cyclic monotonicity. Six combinations have non-negative intervals that include zero, consistent with weak monotonicity but not a positive strict margin. One combination, Fever–Llama, has an interval strictly below zero, indicating a weak-monotonicity violation. Domain effects are pronounced: Diabetes performs weakest, with no model showing a strictly positive interval and an average signed margin near zero. Across models, DeepSeek has the largest average signed margin.

Prompt Consistency. In Table 3 (Appendix), we report the standard deviation of elicited beliefs under the standard prompt and the RMSE deviation between elicited beliefs under the standard prompt and an alternative MSE-framed prompt. We find that for GPT-Min, GPT-High, and DeepSeek, the MSE-framed prompt produces deviations that are comparable to within-prompt variability under repeated standard prompting. For Llama, however, the MSE framing induces substantially larger differences in elicited probabilities. Overall, GPT-High displays the lowest cross-prompt variability, while Llama exhibits the largest sensitivity to the alternative task framing.

Internal Consistency Test. As in Section 3.2, we elicit (i) the model's marginal belief $p(x)$ that a patient has the condition and (ii) the beliefs needed for the LIE decomposition over an auxiliary "next-state" variable z. We measure probabilistic coherence using the Law of Iterated Expectation error $\Delta_{\text{LIE}}(x)$, summarized in Figure 1 by the median normalized ratio $\Delta_{\text{LIE}}(x)/p(x)$ for robustness when $p(x)$ is near zero. As a baseline, we compute the analogous ratio using cross-validated random forest predictors trained on ground-truth $p^*(x)$ and $p^*(x, z)$. Across datasets, LLMs typically show substantially larger inconsistency than this baseline, with no model consistently dominating. This suggests that probabilistic coherence is distinct from decision-consistency.

Robustness. Proposition 1 holds for every utility $u(a, \theta)$ and regularizer C, so no preference over actions and outcomes, however asymmetric, can produce a conditional-independence violation. Only a utility that varies with the evidence x is excluded, since it could rationalize arbitrary actions without beliefs about θ. Empirically, such preferences do not drive our result. Sufficiency violations remain when we run the test separately within strata of each evidence covariate and when we write the payoff matrix directly into the decision prompt, so that the model's utility is given rather than inferred (Appendix H). The result is also not an estimation artifact: it holds across different neighbour counts in the CMI estimator, under a discrete plug-in estimator, under a permutation null that shuffles actions only among cases with similar reported beliefs, when we resample whole contexts rather than individual rows, and when beliefs come from alternative elicitation prompts. We also find the same conclusions in an exploratory extension with a fourth action, ordering a diagnostic test (Appendix H).

Discussion and Conclusions. We develop a novel black-box decision-theoretic framework for testing whether elicited LLM probabilities act as subjective beliefs guiding decisions. Because the framework allows unrestricted utilities and stochastic choice, test failures indicate not merely deviations from expected utility, but failures of elicited probabilities to summarize decision-relevant information. Theoretically, Proposition 5 shows that conditional independence and monotonicity are necessary and sufficient for the observed distribution to be rationalized by a perturbed utility maximizer using the elicited belief as its subjective belief. Empirically, across medical diagnosis tasks and LLMs, elicited beliefs often fail the sufficiency test, although choices frequently remain monotone in elicited risk. For the strongest models, however, these failures are small: elicited probabilities are imperfect but behaviorally informative summaries of decision-relevant information. Thus, elicited probabilities can help interpret model decisions, but should be empirically validated rather than assumed to faithfully represent model beliefs.

AI USE STATEMENT

In this work, we used generative AI tools to help implement and run the experimental pipeline (writing and refactoring analysis and plotting code) and for light editing of the text (grammar, spelling, and word choice). We have not used generative AI tools to generate research ideas, to derive the theoretical results, or to write the substantive content of the paper, and the remaining disclosure categories are not applicable to this work. Large language models are also the objects of study; the model outputs used as data are fully specified by the prompts in Appendix B. We have reviewed all AI-assisted work: AI-generated code was checked against independent implementations and its outputs were verified by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All theoretical claims are stated in Section 3 with complete proofs in Appendix A. Every prompt template, the evidence-to-language conversion, and the clinical question for each dataset are given verbatim in Appendix B; the construction of the two public datasets is described in Appendix D and Appendix E, and estimator settings in Appendix F. Robustness of the main tests to the estimator, the null construction, stated payoffs, and the elicitation prompt is reported in Appendix H. Code implementing the tests, the bootstrap and permutation procedures, and the figure generation, along with the elicited beliefs and decisions, will be released with the paper.

REFERENCES

Ankur Ankan and Johannes Textor. pgmpy: A python toolkit for bayesian networks. *Journal of Machine Learning Research*, 25(265):1–8, 2024. URL http://jmlr.org/papers/v25/23-0487.html.

Francis Anscombe and Robert Aumann. A definition of subjective probability. *The Annals of Mathematical Statistics*, 34, 03 1963. doi: 10.1214/aoms/1177704255.

Amos Azaria and Tom Mitchell. The internal state of an llm knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976, 2023.

Leonard Bereska and Efstratios Gavves. Mechanistic interpretability for ai safety–a review, 2024. *URL https://arxiv. org/abs/2404.14082*.

Marcel Binz and Eric Schulz. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120, 2023.

Mariana Blanco, Dirk Engelmann, Alexander K Koch, and Hans-Theo Normann. Belief elicitation in experiments: is there a hedging problem? *Experimental economics*, 13(4):412–438, 2010.

CDC. Cdc diabetes health indicators, 2017. URL https://archive.ics.uci.edu/dataset/891.

Gary Charness, Uri Gneezy, and Vlastimil Rasocha. Experimental methods: Eliciting beliefs. *Journal of Economic Behavior & Organization*, 189:234–256, 2021.

Vanessa Cheung, Maximilian Maier, and Falk Lieder. Large language models show amplified cognitive biases in moral decision-making. *Proceedings of the National Academy of Sciences*, 122 (25):e2412015122, 2025.

Rachel TA Croson. Thinking like a game theorist: factors affecting the frequency of equilibrium play. *Journal of economic behavior & organization*, 41(3):299–314, 2000.

André F Cruz, Moritz Hardt, and Celestine Mendler-Dünner. Evaluating language models as risk scores. *Advances in Neural Information Processing Systems*, 37:97378–97407, 2024.

Jessica Maria Echterhoff, Yao Liu, Abeer Alessa, Julian McAuley, and Zexue He. Cognitive bias in decision-making with llms. In *Findings of the association for computational linguistics: EMNLP 2024*, pp. 12640–12653, 2024.

Pierre Elias and Joshua Finer. Echonext: A dataset for detecting echocardiogram-confirmed structural heart disease from ecgs, Sep 2025. URL https://physionet.org/content/echonext/1.1.0/.

Mogens Fosgerau, Emerson Melo, Andre De Palma, and Matthew Shum. Discrete choice and rational inattention: A general equivalence result. *International economic review*, 61(4):1569–1589, 2020.

Gabriel Freedman and Francesca Toni. Exploring the potential for large language models to demonstrate rational probabilistic beliefs. *arXiv preprint arXiv:2504.13644*, 2025.

Drew Fudenberg, Ryota Iijima, and Tomasz Strzalecki. Stochastic choice and revealed perturbed utility. *Econometrica*, 83(6):2371–2409, 2015.

Farieda Gaber, Maqsood Shaik, Fabio Allega, Agnes Julia Bilecz, Felix Busch, Kelsey Goon, Vedran Franke, and Altuna Akalin. Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis, May 2025. URL https://www.nature.com/articles/s41746-025-01684-1.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Maxime Griot, Coralie Hemptinne, Jean Vanderdonckt, and Demet Yuksel. Large language models lack essential metacognition for reliable medical reasoning. *Nature Communications*, 16, 01 2025. doi: 10.1038/s41467-024-55628-6.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645 (8081):633–638, September 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL http://dx.doi.org/10.1038/s41586-025-09422-z.

Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, and Daniel Rueckert. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine*, 30:2613–2622, 07 2024. doi: 10.1038/s41591-024-03097-1.

Daniel A Herrmann and Benjamin A Levinstein. Standards for belief representations in llms. *Minds and Machines*, 35(1):5, 2024.

Josef Hofbauer and William H Sandholm. On the global convergence of stochastic fictitious play. *Econometrica*, 70(6):2265–2294, 2002.

Shreya Johri, Jaehwan Jeong, Benjamin A Tran, Daniel I Schlessinger, Shannon Wongvibulsin, Leandra A Barnes, Hong-Yu Zhou, Zhuo Ran Cai, Eliezer M Van Allen, David Kim, et al. An evaluation framework for clinical use of large language models in patient interaction tasks. *Nature medicine*, 31(1):77–86, 2025.

Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291, 1979. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1914185.

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

Ryan Liu, Jiayi Geng, Joshua C Peterson, Ilia Sucholutsky, and Thomas L Griffiths. Large language models assume people are more rational than we really are. *arXiv preprint arXiv:2406.17055*, 2024.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*, 2023.

Daniel McFadden. Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics*, pp. 105–142, 1973.

Daniel McFadden and Marcel K Richter. Stochastic rationality and revealed stochastic preference. *Preferences, Uncertainty, and Optimality, Essays in Honor of Leo Hurwicz, Westview Press: Boulder, CO*, pp. 161–186, 1990.

MetaAI. Introducing llama 4: Advancing multimodal intelligence. https://ai.meta.com/blog/llama-4-multimodal-intelligence/, 2024.

Yaw Nyarko and Andrew Schotter. An experimental study of belief learning using elicited beliefs. *Econometrica*, 70(3):971–1005, 2002.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

Arka Pal, Teo Kitanovski, Arthur Liang, Akilesh Potti, and Micah Goldblum. Incoherent beliefs & inconsistent actions in large language models, 2025. URL https://arxiv.org/abs/2511.13240.

Pedro Rey-Biel. Equilibrium play and best response to (stated) beliefs in normal form games. *Games and Economic Behavior*, 65(2):572–585, 2009.

R Tyrrell Rockafellar. *Convex analysis*, volume 28. Princeton university press, 1997.

David Ronayne, Roberto Veneziani, and William R Zame. Do decision makers have subjective probabilities? an experimental test. *SSRN*, 2022.

Leonard J Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971.

Leonard J Savage. *The foundations of statistics*. Courier Corporation, 1972.

Karl H Schlag, James Tremewan, and Joël J Van der Weele. A penny for your thoughts: A survey of methods for eliciting beliefs. *Experimental Economics*, 18(3):457–490, 2015.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, et al. Open problems in mechanistic interpretability. *arXiv preprint arXiv:2501.16496*, 2025.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, Alex Makelov, Alex Neitz, Alex Wei, Alexandra Barr, Alexandre Kirchmeyer, Alexey Ivanov, Alexi Christakis, Alistair Gillespie, Allison Tam, Ally Bennett, Alvin Wan, Alyssa Huang, Amy McDonald Sandjideh, Amy Yang, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrei Gheorghe, Andres Garcia Garcia, Andrew Braunstein, Andrew Liu, Andrew Schmidt, Andrey Mereskin, Andrey Mishchenko, Andy Applebaum, Andy Rogerson, Ann Rajan, Annie Wei, Anoop Kotha, Anubha Srivastava, Anushree Agrawal, Arun Vijayvergiya, Ashley Tyra, Ashvin Nair, Avi Nayak, Ben Eggers, Bessie Ji, Beth Hoover, Bill Chen, Blair Chen, Boaz Barak, Borys Minaiev, Botao Hao, Bowen Baker, Brad Lightcap, Brandon McKinzie, Brandon Wang, Brendan Quinn, Brian Fioca, Brian Hsu, Brian Yang, Brian Yu, Brian Zhang, Brittany Brenner, Callie Riggins Zetino, Cameron Raymond, Camillo Lugaresi, Carolina Paz, Cary Hudson, Cedric Whitney, Chak Li, Charles Chen, Charlotte Cole, Chelsea Voss, Chen Ding, Chen Shen, Chengdu Huang, Chris Colby, Chris Hallacy, Chris Koch, Chris Lu, Christina Kaplan, Christina Kim, CJ Minott-Henriques, Cliff Frey, Cody Yu, Coley Czarnecki, Colin Reid, Colin Wei, Cory Decareaux, Cristina Scheau, Cyril Zhang, Cyrus Forbes, Da Tang, Dakota Goldberg, Dan Roberts, Dana Palmie, Daniel Kappler, Daniel Levine, Daniel Wright, Dave Leo, David Lin, David Robinson, Declan Grabb, Derek Chen, Derek Lim, Derek Salama, Dibya Bhattacharjee, Dimitris Tsipras, Dinghua Li, Dingli Yu, DJ Strouse, Drew Williams, Dylan Hunn, Ed Bayes, Edwin Arbus, Ekin

Akyurek, Elaine Ya Le, Elana Widmann, Eli Yani, Elizabeth Proehl, Enis Sert, Enoch Cheung, Eri Schwartz, Eric Han, Eric Jiang, Eric Mitchell, Eric Sigler, Eric Wallace, Erik Ritter, Erin Kavanaugh, Evan Mays, Evgenii Nikishin, Fangyuan Li, Felipe Petroski Such, Filipe de Avila Belbute Peres, Filippo Raso, Florent Bekerman, Foivos Tsimpourlas, Fotis Chantzis, Francis Song, Francis Zhang, Gaby Raila, Garrett McGrath, Gary Briggs, Gary Yang, Giambattista Parascandolo, Gildas Chabot, Grace Kim, Grace Zhao, Gregory Valiant, Guillaume Leclerc, Hadi Salman, Hanson Wang, Hao Sheng, Haoming Jiang, Haoyu Wang, Haozhun Jin, Harshit Sikchi, Heather Schmidt, Henry Aspegren, Honglin Chen, Huida Qiu, Hunter Lightman, Ian Covert, Ian Kivlichan, Ian Silber, Ian Sohl, Ibrahim Hammoud, Ignasi Clavera, Ikai Lan, Ilge Akkaya, Ilya Kostrikov, Irina Kofman, Isak Etinger, Ishaan Singal, Jackie Hehir, Jacob Huh, Jacqueline Pan, Jake Wilczynski, Jakub Pachocki, James Lee, James Quinn, Jamie Kiros, Janvi Kalra, Jasmyn Samaroo, Jason Wang, Jason Wolfe, Jay Chen, Jay Wang, Jean Harb, Jeffrey Han, Jeffrey Wang, Jennifer Zhao, Jeremy Chen, Jerene Yang, Jerry Tworek, Jesse Chand, Jessica Landon, Jessica Liang, Ji Lin, Jiancheng Liu, Jianfeng Wang, Jie Tang, Jihan Yin, Joanne Jang, Joel Morris, Joey Flynn, Johannes Ferstad, Johannes Heidecke, John Fishbein, John Hallman, Jonah Grant, Jonathan Chien, Jonathan Gordon, Jongsoo Park, Jordan Liss, Jos Kraaijeveld, Joseph Guay, Joseph Mo, Josh Lawson, Josh McGrath, Joshua Vendrow, Joy Jiao, Julian Lee, Julie Steele, Julie Wang, Junhua Mao, Kai Chen, Kai Hayashi, Kai Xiao, Kamyar Salahi, Kan Wu, Karan Sekhri, Karan Sharma, Karan Singhal, Karen Li, Kenny Nguyen, Keren Gu-Lemberg, Kevin King, Kevin Liu, Kevin Stone, Kevin Yu, Kristen Ying, Kristian Georgiev, Kristie Lim, Kushal Tirumala, Kyle Miller, Lama Ahmad, Larry Lv, Laura Clare, Laurance Fauconnet, Lauren Itow, Lauren Yang, Laurentia Romaniuk, Leah Anise, Lee Byron, Leher Pathak, Leon Maksin, Leyan Lo, Leyton Ho, Li Jing, Liang Wu, Liang Xiong, Lien Mamitsuka, Lin Yang, Lindsay McCallum, Lindsey Held, Liz Bourgeois, Logan Engstrom, Lorenz Kuhn, Louis Feuvrier, Lu Zhang, Lucas Switzer, Lukas Kondraciuk, Lukasz Kaiser, Manas Joglekar, Mandeep Singh, Mandip Shah, Manuka Stratta, Marcus Williams, Mark Chen, Mark Sun, Marselus Cayton, Martin Li, Marvin Zhang, Marwan Aljubeh, Matt Nichols, Matthew Haines, Max Schwarzer, Mayank Gupta, Meghan Shah, Melody Huang, Meng Dong, Mengqing Wang, Mia Glaese, Micah Carroll, Michael Lampe, Michael Malek, Michael Sharman, Michael Zhang, Michele Wang, Michelle Pokrass, Mihai Florian, Mikhail Pavlov, Miles Wang, Ming Chen, Mingxuan Wang, Minnia Feng, Mo Bavarian, Molly Lin, Moose Abdool, Mostafa Rohaninejad, Nacho Soto, Natalie Staudacher, Natan LaFontaine, Nathan Marwell, Nelson Liu, Nick Preston, Nick Turley, Nicklas Ansman, Nicole Blades, Nikil Pancha, Nikita Mikhaylin, Niko Felix, Nikunj Handa, Nishant Rai, Nitish Keskar, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Oona Gleeson, Pamela Mishkin, Patryk Lesiewicz, Paul Baltescu, Pavel Belov, Peter Zhokhov, Philip Pronin, Phillip Guo, Phoebe Thacker, Qi Liu, Qiming Yuan, Qinghua Liu, Rachel Dias, Rachel Puckett, Rahul Arora, Ravi Teja Mullapudi, Raz Gaon, Reah Miyara, Rennie Song, Rishabh Aggarwal, RJ Marsan, Robel Yemiru, Robert Xiong, Rohan Kshirsagar, Rohan Nuttall, Roman Tsiupa, Ronen Eldan, Rose Wang, Roshan James, Roy Ziv, Rui Shu, Ruslan Nigmatullin, Saachi Jain, Saam Talaie, Sam Altman, Sam Arnesen, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Sarah Yoo, Savannah Heon, Scott Ethersmith, Sean Grove, Sean Taylor, Sebastien Bubeck, Sever Banesiu, Shaokyi Amdo, Shengjia Zhao, Sherwin Wu, Shibani Santurkar, Shiyu Zhao, Shraman Ray Chaudhuri, Shreyas Krishnaswamy, Shuaiqi, Xia, Shuyang Cheng, Shyamal Anadkat, Simón Posada Fishman, Simon Tobin, Siyuan Fu, Somay Jain, Song Mei, Sonya Egoian, Spencer Kim, Spug Golden, SQ Mah, Steph Lin, Stephen Imm, Steve Sharpe, Steve Yadlowsky, Sulman Choudhry, Sungwon Eum, Suvansh Sanjeev, Tabarak Khan, Tal Stramer, Tao Wang, Tao Xin, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Degry, Thomas Shadwell, Tianfu Fu, Tianshi Gao, Timur Garipov, Tina Sriskandarajah, Toki Sherbakov, Tomer Kaftan, Tomo Hiratsuka, Tongzhou Wang, Tony Song, Tony Zhao, Troy Peterson, Val Kharitonov, Victoria Chernova, Vineet Kosaraju, Vishal Kuo, Vitchyr Pong, Vivek Verma, Vlad Petrov, Wanning Jiang, Weixing Zhang, Wenda Zhou, Wenlei Xie, Wenting Zhan, Wes McCabe, Will DePue, Will Ellsworth, Wulfie Bain, Wyatt Thompson, Xiangning Chen, Xiangyu Qi, Xin Xiang, Xinwei Shi, Yann Dubois, Yaodong Yu, Yara Khakbaz, Yifan Wu, Yilei Qian, Yin Tat Lee, Yinbo Chen, Yizhen Zhang, Yizhong Xiong, Yonglong Tian, Young Cha, Yu Bai, Yu Yang, Yuan Yuan, Yuanzhi Li, Yufeng Zhang, Yuguang Yang, Yujia Jin, Yun Jiang, Yunyun Wang, Yushi Wang, Yutian Liu, Zach Stubenvoll, Zehao Dou, Zheng Wu, and Zhigang Wang. Openai gpt-5 system card, 2025. URL https://arxiv.org/abs/2601.03267.

K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, M. Amin, L. Hou, K. Clark, S. Pfohl, H. Cole-Lewis, D. Neal, Q. Rashid, M. Schaekermann, A. Wang, D. Dash, J. Chen, N. Shah,

S. Lachgar, P. Mansfield, S. Prakash, B. Green, E. Dominowska, B. Agüera y Arcas, N. Tomašev, Y. Liu, R. Wong, C. Semturs, S. Mahdavi, J. Barral, D. Webster, G. Corrado, Y. Matias, S. Azizi, A. Karthikesalingam, and V. Natarajan. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31:943–950, January 2025. doi: 10.1038/s41591-024-03423-7. URL https://doi.org/10.1038/s41591-024-03423-7.

Kenneth E Train. *Discrete Choice Methods with Simulation*. Cambridge University Press, 2nd edition, 2009.

Dennis Ulmer, Martin Gubri, Hwaran Lee, Sangdoo Yun, and Seong Oh. Calibrating large language models using their generations only. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15440–15459, 2024.

Bo Waggoner and Yiling Chen. Output agreement mechanisms and common knowledge. In *Second AAAI Conference on Human Computation and Crowdsourcing*, HCOMP, 2014.

Cheng Wang, Gyuri Szarvas, Georges Balazs, Pavel Danchenko, and Patrick Ernst. Calibrating verbalized probabilities for large language models. *arXiv preprint arXiv:2410.06707*, 2024.

Christopher Y. K. Williams, Jaskaran Bains, Tianyu Tang, Kishan Patel, Alexa N. Lucas, Fiona Chen, Brenda Y. Miao, Atul J. Butte, and Aaron E. Kornblith. Evaluating large language models for drafting emergency department encounter summaries. *PLOS Digital Health*, 4(6):1–14, 06 2025. doi: 10.1371/journal.pdig.0000899. URL https://doi.org/10.1371/journal.pdig.0000899.

Jian-Qiao Zhu and Thomas L Griffiths. Eliciting the priors of large language models using iterated in-context learning. *arXiv preprint arXiv:2406.01860*, 2024.

Jian-Qiao Zhu and Thomas L. Griffiths. Incoherent probability judgments in large language models, 2025. URL https://arxiv.org/abs/2401.16646.

APPENDIX

A PROOFS

A.1 PROOF OF PROPOSITION 1

Proof. Let $B := P_S(\theta \mid x)$ denote the agent's subjective posterior belief, viewed as a random element of $\Delta(\Theta)$. For each belief value b, define the expected-utility index

$$v_a(b) := \sum_{\theta \in \Theta} b_\theta u(a, \theta).$$

By Definition 1, conditional on $B = b$, the agent's choice-probability vector is

$$q(b) \in \arg\max_{\lambda \in \Delta(\mathcal{A})} \left\{ \sum_{a \in \mathcal{A}} \lambda_a v_a(b) - C(\lambda) \right\}.$$

Thus the conditional distribution of the action depends on the context x only through the subjective belief B.

Fix any action $a \in \mathcal{A}$. Since $\Pr(A = a \mid x, \theta) = q_a(B)$, we have

$$\Pr(A = a \mid B, \theta) = \mathbb{E}[\Pr(A = a \mid x, \theta) \mid B, \theta] = \mathbb{E}[q_a(B) \mid B, \theta] = q_a(B).$$

Similarly, $\Pr(A = a \mid B) = q_a(B)$. Therefore

$$\Pr(A = a \mid B, \theta) = \Pr(A = a \mid B)$$

for every action a, which implies

$$A \perp\!\!\!\perp \theta \mid B.$$

Equivalently,

$$a \perp\!\!\!\perp \theta \mid P_S(\theta \mid x).$$

For the converse statement, suppose the elicited belief is the agent's subjective belief, so that $P_E(\theta \mid x) = P_S(\theta \mid x)$ almost surely. Then the argument above implies

$$a \perp\!\!\!\perp \theta \mid P_E(\theta \mid x).$$

Hence, if the observed joint distribution satisfies

$$a \not\perp\!\!\!\perp \theta \mid P_E(\theta \mid x),$$

then no agent satisfying Definition 1, under any utility function u and regularizer C, can have $P_S = P_E$ while generating the observed actions. □

A.1.1 CONDITIONAL INDEPENDENCE UNDER PROSPECT-THEORETIC RISK AVERSION

Proof. The conditional-independence conclusion does not rely on expected-utility linearity. It only requires that, conditional on the agent's subjective belief, the distribution over actions is fixed.

Let $B := P_S(\theta \mid x)$ denote the agent's subjective posterior belief. Suppose the agent evaluates each action using a prospect-theoretic index $V_a(B)$, where $V_a(B)$ may incorporate a value function, reference dependence, loss aversion, and probability weighting, but depends on the context x only through B. For example, one may write $V_a(B) = \sum_{\theta \in \Theta} \omega_\theta(B) \phi(u(a, \theta) - r_a)$, where ϕ is a prospect-theoretic value function, $\omega(B)$ is a deterministic probability-weighting transformation of the belief B, and r_a is a reference point. More generally, the argument only requires that $V_a(B)$ is measurable with respect to B.

The prospect-theoretic perturbed choice rule is then

$$q(B) \in \arg\max_{\lambda \in \Delta(\mathcal{A})} \left\{ \sum_{a \in \mathcal{A}} \lambda_a V_a(B) - C(\lambda) \right\}.$$

Thus, conditional on B, the choice-probability vector is a deterministic function of B. In particular, for every action $a \in \mathcal{A}$, $\Pr(A = a \mid x, \theta) = q_a(B)$. Therefore,

$$\Pr(A = a \mid B, \theta) = \mathbb{E}[\Pr(A = a \mid x, \theta) \mid B, \theta] = \mathbb{E}[q_a(B) \mid B, \theta] = q_a(B).$$

Similarly, $\Pr(A = a \mid B) = q_a(B)$. Hence $\Pr(A = a \mid B, \theta) = \Pr(A = a \mid B)$ for every action a, which implies $A \perp\!\!\!\perp \theta \mid B$.

Consequently, if the elicited belief equals the subjective belief, so that $P_E(\theta \mid x) = P_S(\theta \mid x)$ almost surely, then $A \perp\!\!\!\perp \theta \mid P_E(\theta \mid x)$ also holds. Thus a violation of this conditional-independence restriction cannot be explained merely by prospect-theoretic risk aversion, provided the prospect-theoretic evaluation depends on the context only through the agent's subjective belief. □

A.2 Proof of Proposition 2

Proof. Let $b_1, \ldots, b_M$ be any finite collection of beliefs, and set $b_{M+1} = b_1$. Write

$$v_m := v(b_m), \qquad q_m := q(b_m).$$

By Definition 1, for each m,

$$q_m \in \arg\max_{\lambda \in \Delta(\mathcal{A})} \{\lambda \cdot v_m - C(\lambda)\}.$$

Therefore, comparing q_m to q_{m-1} at the utility index v_m, with indices interpreted cyclically, we have

$$q_m \cdot v_m - C(q_m) \geq q_{m-1} \cdot v_m - C(q_{m-1}).$$

Rearranging,

$$v_m \cdot (q_m - q_{m-1}) \geq C(q_m) - C(q_{m-1}).$$

Summing over $m = 1, \ldots, M$ gives

$$\sum_{m=1}^{M} v_m \cdot (q_m - q_{m-1}) \geq \sum_{m=1}^{M} \big(C(q_m) - C(q_{m-1})\big).$$

The right-hand side telescopes to zero, so

$$\sum_{m=1}^{M} v_m \cdot (q_m - q_{m-1}) \geq 0.$$

Reindexing the second term yields

$$\sum_{m=1}^{M} q_m \cdot (v_m - v_{m+1}) \geq 0.$$

Substituting back $q_m = q(b_m)$ and $v_m = v(b_m)$ gives

$$\sum_{m=1}^{M} q(b_m) \cdot \big(v(b_m) - v(b_{m+1})\big) \geq 0,$$

which is cyclic monotonicity.

The final claim follows by contrapositive: if the observed conditional choice-probability map violates one of these inequalities, then it cannot be generated by any agent satisfying Definition 1 using the elicited probabilities as the operative belief index. □

A.3 Proof of Proposition 3

Proof. For a binary state, the expected-utility index for action a at belief p is

$$v_a(p) = (1 - p)u(a, 0) + pu(a, 1) = u(a, 0) + pd_a,$$

where $d_a := u(a, 1) - u(a, 0)$. Hence, for any p, p',

$$v(p') - v(p) = (p' - p)d.$$

Define

$$m(p) := q(p) \cdot d = \sum_{a=1}^{J} q_a(p) d_a.$$

We first show that cyclic monotonicity with respect to $v(p)$ is equivalent to weak monotonicity of $m(\cdot)$. Cyclic monotonicity requires that, for every finite cycle $p_1, \ldots, p_M$ with $p_{M+1} = p_1$,

$$\sum_{k=1}^{M} q(p_k) \cdot \big(v(p_k) - v(p_{k+1})\big) \geq 0.$$

Using

$$v(p_k) - v(p_{k+1}) = (p_k - p_{k+1})d,$$

this condition is equivalent to

$$\sum_{k=1}^{M} m(p_k)(p_k - p_{k+1}) \geq 0.$$

We now prove that this one-dimensional cycle condition is equivalent to weak monotonicity of m. First suppose the cycle condition holds. Consider any $p < p'$ and apply the condition to the two-point cycle (p, p'). Then

$$m(p)(p - p') + m(p')(p' - p) \geq 0.$$

Equivalently,

$$(p' - p)\big(m(p') - m(p)\big) \geq 0.$$

Since $p' - p > 0$, this implies

$$m(p) \leq m(p').$$

Thus m is weakly increasing.

Conversely, suppose m is weakly increasing. Let $p_1, \ldots, p_M$ be any finite cycle with $p_{M+1} = p_1$. Since only finitely many belief values appear in the cycle, list their distinct values as

$$r_1 < r_2 < \cdots < r_L.$$

Define a function M on these grid points by setting $M(r_1) = 0$ and, for $i = 1, \ldots, L-1$,

$$M(r_{i+1}) - M(r_i) = m(r_i)(r_{i+1} - r_i).$$

Extend M linearly on each interval $[r_i, r_{i+1}]$. Because m is weakly increasing, the slopes of this piecewise-linear function are weakly increasing. Therefore M is convex.

Moreover, for every grid point r_i, the number $m(r_i)$ is a subgradient of M at r_i. Indeed, the left slope at r_i is weakly below $m(r_i)$ and the right slope at r_i is weakly above $m(r_i)$, with the natural one-sided interpretation at the endpoints. Hence, for every k,

$$M(p_{k+1}) \geq M(p_k) + m(p_k)(p_{k+1} - p_k).$$

Rearranging gives

$$m(p_k)(p_k - p_{k+1}) \geq M(p_k) - M(p_{k+1}).$$

Summing over $k = 1, \ldots, M$ yields

$$\sum_{k=1}^{M} m(p_k)(p_k - p_{k+1}) \geq \sum_{k=1}^{M} \big(M(p_k) - M(p_{k+1})\big) = 0,$$

where the final equality follows from $p_{M+1} = p_1$. Thus the one-dimensional cycle condition holds. Therefore, cyclic monotonicity with respect to $v(p)$ holds if and only if m is weakly increasing.

It follows that cyclic monotonicity with respect to $v(p)$ holds if and only if, for every $p < p'$,

$$m(p) \leq m(p').$$

Expanding the definition of m gives

$$\sum_{a=1}^{J} q_a(p) d_a \leq \sum_{a=1}^{J} q_a(p') d_a.$$

This proves the equivalence claimed in the first part of the proposition.

It remains to prove the strict part under Definition 1. Suppose the agent satisfies Definition 1. Let

$$q := q(p), \qquad q' := q(p').$$

Optimality of q at $v(p)$ and of q' at $v(p')$ implies

$$q \cdot v(p) - C(q) \geq q' \cdot v(p) - C(q')$$

and

$$q' \cdot v(p') - C(q') \geq q \cdot v(p') - C(q).$$

Adding these two inequalities gives

$$q \cdot v(p) + q' \cdot v(p') \geq q' \cdot v(p) + q \cdot v(p').$$

Rearranging,

$$(q' - q) \cdot \big(v(p') - v(p)\big) \geq 0.$$

Since

$$v(p') - v(p) = (p' - p)d$$

and $p' - p > 0$, this implies

$$(q' - q) \cdot d \geq 0,$$

or equivalently

$$q' \cdot d \geq q \cdot d.$$

Now suppose $q \neq q'$. For each fixed v, the objective

$$\lambda \mapsto \lambda \cdot v - C(\lambda)$$

is strictly concave on $\Delta(\mathcal{A})$ because C is strictly convex. Therefore the optimizer at each fixed v is unique. Since q is the optimizer at $v(p)$ and $q' \neq q$, the first optimality inequality is strict:

$$q \cdot v(p) - C(q) > q' \cdot v(p) - C(q').$$

Similarly, since q' is the optimizer at $v(p')$ and $q \neq q'$, the second optimality inequality is strict:

$$q' \cdot v(p') - C(q') > q \cdot v(p') - C(q).$$

Adding the strict inequalities yields

$$(q' - q) \cdot \big(v(p') - v(p)\big) > 0.$$

Again using $v(p') - v(p) = (p' - p)d$ and $p' - p > 0$, we obtain

$$q' \cdot d > q \cdot d.$$

Thus the monotonicity inequality is strict whenever $q(p) \neq q(p')$. □

A.4 Proof of Proposition 4

Proof. We first prove the strict part. Suppose there exists a non-constant vector $d' \in \mathbb{R}^J$ such that $\widehat{m}_1(d') \leq \cdots \leq \widehat{m}_K(d')$ and $\widehat{m}_k(d') < \widehat{m}_{k+1}(d')$ for every $k \in \mathcal{S}$. Since $\mathcal{S}$ is finite and nonempty,

$$\delta := \min_{k \in \mathcal{S}} \{\widehat{m}_{k+1}(d') - \widehat{m}_k(d')\} > 0.$$

Let $\alpha := \min_a d'_a$ and $\beta := \max_a d'_a$. Since d' is non-constant, $\beta > \alpha$. Define

$$\widetilde{d}_a := \frac{d'_a - \alpha}{\beta - \alpha}.$$

Then $\widetilde{d} \in [0,1]^J$, and there exist actions $a^\star, b^\star$ such that $\widetilde{d}_{a^\star} = 0$ and $\widetilde{d}_{b^\star} = 1$. Because each $\widehat{q}_k$ is a probability vector,

$$\widehat{m}_k(\widetilde{d}) = \frac{\widehat{m}_k(d') - \alpha}{\beta - \alpha}.$$

Thus the affine rescaling preserves all weak and strict inequalities, and the minimum increment over $\mathcal{S}$ becomes $\delta/(\beta-\alpha) > 0$. Therefore $(\widetilde{d}, \gamma)$ is feasible for $\mathrm{SignedLP}(a^\star, b^\star)$ for any $0 < \gamma \leq \delta/(\beta-\alpha)$, so $\widehat{\gamma} > 0$.

Conversely, suppose $\widehat{\gamma} > 0$. Then for some ordered pair $(a^\star, b^\star)$ there exists a feasible solution (d, γ) to $\mathrm{SignedLP}(a^\star, b^\star)$ with $\gamma > 0$. Hence

$$\widehat{m}_{k+1}(d) - \widehat{m}_k(d) \geq \gamma > 0, \qquad k \in \mathcal{S}.$$

For $k \notin \mathcal{S}$, we have $\widehat{q}_k = \widehat{q}_{k+1}$, so $\widehat{m}_{k+1}(d) - \widehat{m}_k(d) = 0$. Therefore $\widehat{m}_1(d) \leq \cdots \leq \widehat{m}_K(d)$, with strict inequality for every $k \in \mathcal{S}$. Moreover, d is non-constant because $d_{a^\star} = 0$ and $d_{b^\star} = 1$.

The weak part is analogous. If there exists a non-constant $d' \in \mathbb{R}^J$ such that $\widehat{m}_1(d') \leq \cdots \leq \widehat{m}_K(d')$, then the same affine rescaling produces a normalized vector $\widetilde{d} \in [0,1]^J$ attaining both 0 and 1. For the corresponding ordered pair $(a^\star, b^\star)$,

$$\widehat{m}_{k+1}(\widetilde{d}) - \widehat{m}_k(\widetilde{d}) \geq 0, \qquad k \in \mathcal{S},$$

so $(\widetilde{d}, 0)$ is feasible for $\mathrm{SignedLP}(a^\star, b^\star)$ and hence $\widehat{\gamma} \geq 0$. Conversely, if $\widehat{\gamma} \geq 0$, then some feasible solution satisfies $\gamma \geq 0$, which implies weakly nonnegative increments on $\mathcal{S}$. The increments outside $\mathcal{S}$ are identically zero, so $\widehat{m}_1(d) \leq \cdots \leq \widehat{m}_K(d)$. Again, d is non-constant because it attains both 0 and 1.

It remains to show that $\widehat{\gamma} \in [-1, 1]$. For any feasible $d \in [0,1]^J$, each $\widehat{m}_k(d)$ lies in $[0,1]$ because $\widehat{q}_k$ is a probability vector. Hence every adjacent increment $\widehat{m}_{k+1}(d) - \widehat{m}_k(d)$ lies in $[-1,1]$, so any feasible γ satisfying the signed-margin constraints must satisfy $\gamma \leq 1$. For the lower bound, each signed LP is feasible with $\gamma = -1$, since all adjacent increments are at least -1. Therefore each LP value lies in $[-1,1]$, and taking the maximum over ordered pairs gives $\widehat{\gamma} \in [-1,1]$. □

A.5 Proof of Proposition 5

Proof. Let $B := P_E(\theta \mid x)$ denote the elicited belief, viewed as a random element of $\Delta(\Theta)$. For each belief value b, define $q(b) := \big(\Pr_{\mathbb{P}}(a = a' \mid B = b)\big)_{a' \in \mathcal{A}}$. For a utility function $u : \mathcal{A} \times \Theta \to \mathbb{R}$, define the expected-utility index $v(b) \in \mathbb{R}^{|\mathcal{A}|}$ by $v_a(b) := \sum_{\theta \in \Theta} b_\theta u(a, \theta)$.

We prove both directions.

First suppose there exists an agent satisfying Definition 1, with the strict-convexity requirement on C relaxed to convexity, whose subjective belief satisfies $P_S = B$ and whose induced distribution over (B, θ, a) equals $\mathbb{P}$. The proof of Proposition 1 uses only that the agent's choice probabilities depend on x through its subjective belief, not strict convexity of C. Therefore Proposition 1 applies unchanged and implies $a \perp\!\!\!\perp \theta \mid B$. The proof of Proposition 2 uses only convexity of C, not strict convexity, so it also applies unchanged under this relaxed version of Definition 1. Therefore Proposition 2 implies that $q(B)$ is cyclically monotone with respect to the utility index $v(B)$. Thus the two stated conditions are necessary.

Conversely, suppose $a \perp\!\!\!\perp \theta \mid B$ under the observed distribution $\mathbb{P}$, and suppose that $q(B)$ is cyclically monotone with respect to $v_a(B) = \sum_{\theta \in \Theta} B_\theta u(a, \theta)$ for some nonzero utility function $u : \mathcal{A} \times \Theta \to \mathbb{R}$.

We first construct a convex-perturbation representation of the conditional choice rule q. By Rockafellar's cyclic-monotonicity theorem (Rockafellar, 1997), cyclic monotonicity of the relation $\{(v(B), q(B))\}$ implies that there exists a convex potential $W : \mathbb{R}^{|\mathcal{A}|} \to \mathbb{R} \cup \{+\infty\}$ such that $q(B) \in \partial W(v(B))$ for every elicited belief value B in the support of the observed distribution. By Fenchel duality, $q(B) \in \partial W(v(B))$ if and only if $v(B) \in \partial W^*(q(B))$, where W^* is the convex conjugate of W. Equivalently,

$$q(B) \in \operatorname*{arg\,max}_{\lambda \in \Delta(\mathcal{A})} \{\lambda \cdot v(B) - C(\lambda)\},$$

where C is the restriction of W^* to $\Delta(\mathcal{A})$. The function C is convex, but need not be strictly convex. Therefore the cyclic-monotonicity hypothesis gives the perturbed-maximization representation

$$q(B) \in \operatorname*{arg\,max}_{\lambda \in \Delta(\mathcal{A})} \left\{ \sum_{a \in \mathcal{A}} \lambda_a \left(\sum_{\theta \in \Theta} B_\theta u(a, \theta) \right) - C(\lambda) \right\}.$$

Now construct an agent as follows. The agent faces the same marginal distribution over (B, θ) as in $\mathbb{P}$. Its subjective belief is defined to be $P_S(\theta \mid x) = B = P_E(\theta \mid x)$. Conditional on B, the agent uses the choice-probability vector $q(B)$ obtained above and draws its realized action according to $\Pr_{\text{agent}}(a \mid B) = q_a(B)$. By construction, this agent satisfies Definition 1 with nonzero utility function u and with a convex, but not necessarily strictly convex, regularizer C.

It remains to verify that the constructed agent reproduces the full observed joint distribution over (B, θ, a). For any B, θ, and a,

$$\Pr_{\text{agent}}(B, \theta, a) = \Pr_{\text{agent}}(B, \theta) \Pr_{\text{agent}}(a \mid B, \theta).$$

By construction,

$$\Pr_{\text{agent}}(B, \theta) = \Pr_{\mathbb{P}}(B, \theta).$$

Moreover, the constructed action rule depends only on B, so

$$\Pr_{\text{agent}}(a \mid B, \theta) = \Pr_{\text{agent}}(a \mid B) = q_a(B).$$

By the conditional-independence assumption in the observed distribution,

$$\Pr_{\mathbb{P}}(a \mid B, \theta) = \Pr_{\mathbb{P}}(a \mid B) = q_a(B).$$

Therefore,

$$\Pr_{\text{agent}}(B, \theta, a) = \Pr_{\mathbb{P}}(B, \theta) q_a(B) = \Pr_{\mathbb{P}}(B, \theta) \Pr_{\mathbb{P}}(a \mid B, \theta) = \Pr_{\mathbb{P}}(B, \theta, a).$$

Hence the constructed agent has subjective belief $P_S = P_E$, satisfies the perturbed-maximization representation with nonzero utility u and convex, but not necessarily strictly convex, regularizer C, and induces exactly the observed distribution over $(P_E(\theta \mid x), \theta, a)$.

This proves the desired equivalence. □

B PROMPTS

This section provides the complete prompt templates used in our experiments. All prompts follow a consistent structure where clinical evidence is converted to natural language descriptions. We use placeholder notation: `<patient_description>` denotes the natural language description of patient evidence (e.g., "is male, is in the 50-64 age group, and QRS duration is prolonged"), and `<clinical_question>` denotes the condition-specific question (e.g., "have moderate or greater structural heart disease").

B.1 PROBABILITY ELICITATION PROMPTS

B.1.1 STANDARD PROBABILITY ELICITATION (PROMPT π_0)

The standard prompt elicits $P_E(\theta \mid x)$ without any additional instructions about scoring rules or reasoning strategies.

```
There is a patient who <patient_description>.  What is
the probability that they <clinical_question>?  Return
probabilities for:  No and Yes.
Respond with exactly 2 lines, one per label, and no extra
text.
Each line MUST be exactly:  '<label>:  <number between 0 and
1>'.
Use these labels in this order:  [No, Yes]
Example:
No:  0.50
Yes:  0.50
```

B.1.2 MSE SCORING RULE PROMPT (π_{MSE})

This prompt includes an explicit statement that responses will be evaluated under the Mean Squared Error (Brier) scoring rule, which is strictly proper.

```
IMPORTANT: Your probability estimates will be evaluated
according to the Mean Squared Error (MSE) scoring rule.
This means the loss is calculated as the squared difference
between your probability estimate and the true outcome.
Please provide your best probability estimate.
There is a patient who <patient_description>.  What is
the probability that they <clinical_question>?  Return
probabilities for:  No and Yes.
Respond with exactly 2 lines, one per label, and no extra
text.
Each line MUST be exactly:  '<label>:  <number between 0 and
1>'.
Use these labels in this order:  [No, Yes]
Example:
No:  0.50
Yes:  0.50
```

B.1.3 ABSOLUTE LOSS SCORING RULE PROMPT (π_{ABS})

This prompt specifies evaluation under absolute loss.

```
IMPORTANT: Your probability estimates will be evaluated
according to the Absolute Loss scoring rule.  This means the
loss is calculated as the absolute difference between your
probability estimate and the true outcome.  Please provide
your best probability estimate.
There is a patient who <patient_description>.  What is
the probability that they <clinical_question>?  Return
probabilities for:  No and Yes.
```

```
Respond with exactly 2 lines, one per label, and no extra
text.
Each line MUST be exactly: '<label>: <number between 0 and
1>'.
Use these labels in this order: [No, Yes]
Example:
No: 0.50
Yes: 0.50
```

B.1.4 BAYESIAN REASONING PROMPT (π_{BAYES})

This prompt explicitly instructs the model to apply Bayesian reasoning when forming probability estimates.

```
IMPORTANT: Use proper Bayesian reasoning to reach your
answer. Consider the prior probability of the condition
in the general population, then update this prior based on
the evidence.
There is a patient who <patient_description>. What is
the probability that they <clinical_question>? Return
probabilities for: No and Yes.
Respond with exactly 2 lines, one per label, and no extra
text.
Each line MUST be exactly: '<label>: <number between 0 and
1>'.
Use these labels in this order: [No, Yes]
Example:
No: 0.50
Yes: 0.50
```

B.2 DECISION ELICITATION PROMPTS

B.2.1 DECISION PROMPT A: WITHOUT LOSS FUNCTION

This prompt asks the model to make a diagnostic decision without specifying an explicit loss function. The model is asked both (1) whether it feels confident enough to make a decision (allowing for deferral) and (2) what decision it would make if forced to choose.

```
There is a patient who <patient_description>.
Question: Do they <clinical_question>?
Please respond with EXACTLY two lines:
Line 1: Can you make a decision about whether or not the
patient is likely to <clinical_question>? Answer 'Yes' or
'No'.
Line 2: If you had to make a decision, what would it be?
Answer 'Yes' (the patient is likely to <clinical_question>)
or 'No' (the patient is likely to not <clinical_question>)
Example format:
Can decide: No
Decision: Yes
```

B.3 INTERNAL CONSISTENCY (LAW OF ITERATED EXPECTATION) PROMPTS

For the internal consistency test described in Section 3.2, we elicit three sets of probabilities. The base probability $P_E(\theta \mid x)$ uses the standard probability elicitation prompt above. The additional prompts are:

B.3.1 NEXT-STATE DISTRIBUTION PROMPT: $P_E(z \in B_j \mid x)$

This prompt elicits the model's belief about an auxiliary covariate z given the base evidence x.

```
There is a patient who <patient_description>.
What is the probability distribution over their
<auxiliary_variable_name>? Return probabilities for each
category of <auxiliary_variable_name>.
Respond with exactly <K> lines, one per label, and no extra
text.
Each line MUST be exactly: '<label>: <number between 0 and
1>'.
Use these labels in this order: [<state_1>, <state_2>, ...,
<state_K>]
Example:
<state_1>: <1/K>
<state_2>: <1/K>
...
```

B.3.2 CONDITIONAL PROBABILITY PROMPT: $P_E(\theta \mid x, z \in B_j)$

This prompt elicits the model's belief in the target outcome θ given both the base evidence x and a specific value of the auxiliary variable z.

```
There is a patient who <patient_description> and
<auxiliary_variable_condition>. What is the probability
that they <clinical_question>? Return probabilities for:
No and Yes.
Respond with exactly 2 lines, one per label, and no extra
text.
Each line MUST be exactly: '<label>: <number between 0 and
1>'.
Use these labels in this order: [No, Yes]
Example:
No: 0.50
Yes: 0.50
```

B.4 EVIDENCE-TO-LANGUAGE CONVERSION

Patient evidence is converted to natural language using domain-specific clinical phrasing. For the structural heart disease dataset, examples include:

- **Demographics:** "is male," "is in the 50-64 age group," "is in an inpatient setting," "is Hispanic/Latino"
- **ECG findings:** "QRS duration is prolonged," "QTc is not prolonged," "ST–T abnormalities are present," "ECG shows left ventricular hypertrophy"
- **Echocardiographic indicators:** "LVEF $\leq$ 45%," "LV wall thickness $\geq$ 1.3 cm," "moderate or greater aortic stenosis present"

For the pediatric Bayesian networks (fever and crying), evidence phrases are adapted to the relevant symptom domains (e.g., "presents with jaundice," "shows lethargy," "has feeding difficulties," "abdomen is distended").

For the diabetes dataset, evidence includes lifestyle and clinical indicators (e.g., "exercises regularly," "has high glucose levels," "BMI is in the obese range").

B.5 CLINICAL QUESTIONS BY DATASET

The `<clinical_question>` placeholder is instantiated as follows for each dataset:

- **Structural Heart Disease:** "have moderate or greater structural heart disease"
- **Diabetes:** "have diabetes/pre-diabetes"
- **Fever (Pediatric):** "have a fever meeting the threshold ($\geq$ 99°F oral or $\geq$ 100°F rectal)"
- **Infant Crying:** "have colic"

Table 3: **Prompt-consistency metrics.** Within-prompt standard deviation under the standard prompt (π_0), and pairwise RMSE between elicited probabilities under the MSE-framed prompt and π_0.

	π_0	**RMSE: MSE vs π_0**
Dataset/Model	**Std**	**MSE**
Heart–GPT-Min	.0786	.0915
Heart–GPT-High	.0591	.0456
Heart–Llama	.1592	.2403
Heart–DeepSeek	.0781	.0829
Cry–GPT-Min	.1696	.1454
Cry–GPT-High	.0968	.0808
Cry–Llama	.0986	.2090
Cry–DeepSeek	.0948	.0885
Fever–GPT-Min	.1469	.1242
Fever–GPT-High	.1003	.0675
Fever–Llama	.0880	.1716
Fever–DeepSeek	.0890	.0929
Diab–GPT-Min	.0451	.0724
Diab–GPT-High	.0996	.0860
Diab–Llama	.0404	.2520
Diab–DeepSeek	.1004	.1137
Mean: GPT-Min	**.1100**	**.1084**
Mean: GPT-High	**.0889**	**.0700**
Mean: Llama	**.0965**	**.2182**
Mean: DeepSeek	**.0905**	**.0945**
Overall mean	**.0965**	**.1228**

C Prompting Analysis

C.1 Prompt Consistency

In Table 3, we measure the standard deviation of elicited beliefs under the standard prompt and the RMSE deviation in elicited beliefs under an alternative MSE-framed prompt. The alternative prompt tells the LLM that the scoring rule will be mean squared error.

D HEART DISEASE DATASET

D.0.1 DATA SOURCE

The heart disease dataset is derived from electrocardiogram (ECG) and echocardiogram records from Columbia University Medical Center (Elias & Finer, 2025). The dataset contains over 100,000 patient encounters with paired ECG and echocardiographic measurements. Clinical variable thresholds follow the EchoNext v1.1.0 guidelines.

D.0.2 TARGET VARIABLE

The target variable is **Structural Heart Disease (SHD)**, defined as the presence of moderate-or-greater structural heart disease.

D.0.3 COVARIATE VARIABLES

Demographic Variables (4 variables):

- **Age**: Three categories—< 50, 50–69, ≥ 70 years
- **Sex**: Male, Female
- **Location Setting**: Inpatient, Outpatient, Emergency Department, Procedural
- **Race/Ethnicity**: Asian, Black, Hispanic, White, Other, Unknown

ECG Measurements

- **QRS_Prolonged**: Whether QRS duration is prolonged (Yes/No)
- **QTc_Prolonged**: Whether corrected QT interval is prolonged (Yes/No)
- **STT_Abnormal**: Whether ST–T wave abnormalities are present (Yes/No)
- **ECG_LVH**: Whether ECG shows left ventricular hypertrophy (Yes/No)

ECG Indicated Conditions

- Left ventricular ejection fraction (LVEF) $\leq 45\%$
- Left ventricular wall thickness ≥ 1.3 cm
- Moderate or greater aortic stenosis
- Moderate or greater aortic regurgitation
- Moderate or greater mitral regurgitation
- Moderate or greater tricuspid regurgitation
- Moderate or greater pulmonary regurgitation
- Moderate or greater right ventricular systolic dysfunction
- Moderate or large pericardial effusion
- Pulmonary artery systolic pressure (PASP) ≥ 45 mmHg
- Tricuspid regurgitation maximum velocity (TR $V_{\max}$) ≥ 3.2 m/s

D.0.4 BAYESIAN NETWORK STRUCTURE

The Bayesian network contains 20 nodes and 121 directed edges, organized in a three-tier hierarchical structure:

1. **Root nodes**: The four demographic variables (Age, Sex, Location Setting, Race/Ethnicity) serve as root nodes with no parents.

2. **ECG layer**: The four ECG variables are conditionally dependent on all demographic variables. Additionally, there are dependencies among ECG variables: ECG_LVH influences QTc_Prolonged, STT_Abnormal, and QRS_Prolonged; STT_Abnormal influences QTc_Prolonged and QRS_Prolonged; and QTc_Prolonged influences QRS_Prolonged.
3. **Echocardiographic indicator layer**: Each of the 11 echocardiographic indicator flags depends on all demographic variables and all ECG variables.
4. **Target node**: SHD is a direct child of all 11 echocardiographic indicator flags

D.0.5 Ground-Truth Probability Computation

Conditional probability tables (CPTs) are estimated empirically from the dataset using maximum likelihood estimation. For each combination of parent variable states, we compute the empirical frequency of each child state. To ensure reliable probability estimates, we enforce a minimum support threshold of 100 patients for each covariate configuration. Covariate combinations with fewer than 100 observations are excluded from the evaluation to avoid high-variance probability estimates.

Ground-truth posterior probabilities $P^*(\theta \mid x)$ are computed via variable elimination using the `pgmpy` library (Ankan & Textor, 2024). For the experiments, we sample test cases to achieve approximately uniform coverage across the probability range $[0, 1]$ by stratifying into 20 equal-width bins.

D.0.6 Clinical Phrasing Examples

Evidence is converted to natural language for LLM prompts. Example phrasings include:

- "is male, is in the 50–69 age group, is in an inpatient setting, and QRS duration is prolonged"
- "is female, is in the $\geq$70 age group, is Hispanic/Latino, and ST–T abnormalities are present"
- "is in the emergency department, ECG shows left ventricular hypertrophy, and LVEF $\leq$ 45%"

The clinical question for the target variable is phrased as: "have moderate or greater structural heart disease."

E DIABETES DATASET

E.0.1 DATA SOURCE

The diabetes dataset is derived from the CDC Behavioral Risk Factor Surveillance System (BRFSS), a large-scale telephone survey that collects health-related risk behaviors, chronic health conditions, and use of preventive services from U.S. residents (CDC, 2017). The dataset contains self-reported health indicators, lifestyle factors, and demographic information commonly used in diabetes risk prediction models.

E.0.2 TARGET VARIABLE

The target variable is **Diabetes_binary**, indicating self-reported diabetes or prediabetes status:

- **1**: Prediabetes or diabetes (respondent reports having been told by a doctor that they have diabetes or prediabetes)
- **0**: No diabetes reported

E.0.3 COVARIATE VARIABLES

The dataset contains 21 covariate variables organized into five categories:

Demographic Variables (4 variables):

- **Sex**: Female (0), Male (1)
- **Age**: Four categories—Young adult (approximately 18–34 years), Early middle adulthood (approximately 35–49 years), Late middle adulthood (approximately 50–64 years), Older adult (65+ years)
- **Education**: Six levels—No schooling/kindergarten only (1), Elementary education (2), Some high school (3), High school graduate/GED (4), Some college/technical school (5), College graduate or higher (6)
- **Income**: Four categories—Low ($<$\$25,000), Lower-middle (\$25,000–\$49,999), Higher-middle (\$50,000–\$74,999), High ($\geq$\$75,000)

Cardiometabolic and Comorbidity History (6 variables):

- **HighBP**: History of high blood pressure (Yes/No)
- **HighChol**: History of high cholesterol (Yes/No)
- **CholCheck**: Cholesterol check within the past five years (Yes/No)
- **Stroke**: History of stroke (Yes/No)
- **HeartDiseaseorAttack**: History of coronary heart disease or myocardial infarction (Yes/No)
- **BMI**: Body mass index category—Underweight (BMI < 18.5), Normal ($18.5 \leq$ BMI < 25), Overweight ($25 \leq$ BMI < 30), Obese (BMI ≥ 30)

Lifestyle and Behavioral Variables (5 variables):

- **Smoker**: Smoked at least 100 cigarettes in lifetime (Yes/No)
- **PhysActivity**: Physical activity during the past 30 days (Yes/No)
- **Fruits**: Fruit consumption one or more times per day (Yes/No)
- **Veggies**: Vegetable consumption one or more times per day (Yes/No)
- **HvyAlcoholConsump**: Heavy alcohol consumption (Yes/No)

Healthcare Access Variables (2 variables):

- **AnyHealthcare**: Access to health care coverage (Yes/No)
- **NoDocbcCost**: Unable to see a doctor due to cost in the past 12 months (Yes/No)

Self-Reported Health Status Variables (4 variables):

- **GenHlth**: Self-rated general health—Excellent (1), Very good (2), Good (3), Fair (4), Poor (5)
- **MentHlth**: Days of poor mental health in the past 30 days—<7 days (0), 7–13 days (1), 14–20 days (2), 21+ days (3)
- **PhysHlth**: Days of poor physical health in the past 30 days—<7 days (0), 7–13 days (1), 14–20 days (2), 21+ days (3)
- **DiffWalk**: Serious difficulty walking or climbing stairs (Yes/No)

E.0.4 BAYESIAN NETWORK STRUCTURE

The Bayesian network contains 22 nodes and 77 directed edges. Unlike the hierarchical structure of the heart disease network, the diabetes network exhibits a more complex web of interdependencies reflecting the multifactorial nature of diabetes risk. Key structural features include:

1. **Self-rated health as a hub**: `GenHlth` (self-rated general health) serves as a central hub node with edges to 11 other variables, including direct connections to `Diabetes_binary`, `HighBP`, `HighChol`, `BMI`, `HeartDiseaseorAttack`, `Stroke`, and `DiffWalk`.
2. **Functional status pathway**: `DiffWalk` (difficulty walking) connects to `Diabetes_binary`, `HighBP`, `Smoker`, `Stroke`, and demographic variables.
3. **Cardiovascular cascade**: `HighBP` → `HighChol` → `HeartDiseaseorAttack` → `Stroke` forms a causal chain.
4. **Lifestyle clustering**: Diet variables (`Fruits`, `Veggies`) influence `PhysActivity`, which in turn affects `DiffWalk`.
5. **Socioeconomic pathways**: `Education` → `Income` → `GenHlth` and `NoDocbcCost` → `AnyHealthcare` capture socioeconomic determinants of health.
6. **Diabetes as both cause and effect**: `Diabetes_binary` has incoming edges from `GenHlth`, `DiffWalk`, `Sex`, and `Income`, while also having outgoing edges to `HighBP`, `HighChol`, `BMI`, and `Age` (reflecting that diabetes diagnosis may alter other health behaviors and conditions).

E.0.5 GROUND-TRUTH PROBABILITY COMPUTATION

Conditional probability tables are estimated empirically from the BRFSS survey data using maximum likelihood estimation. We enforce a minimum support threshold of 100 respondents for each covariate configuration to ensure reliable probability estimates. Ground-truth posterior probabilities $P^*(\theta \mid x)$ are computed via variable elimination using the `pgmpy` library (Ankan & Textor, 2024).

E.0.6 CLINICAL PHRASING EXAMPLES

Evidence is converted to natural language for LLM prompts. Example phrasings include:

- "is Male, is in the late middle adulthood age group (approximately 50–64 years), has a history of high blood pressure, and is Obese (BMI $\geq$ 30)"
- "is Female, has household income below $25,000, has no daily fruit consumption, and self-rated general health is fair"
- "has access to health care coverage, has physical activity during the past 30 days, and has no history of stroke"

The clinical question for the target variable is phrased as: "have prediabetes or diabetes."

F PARAMETER DETAILS

For Random Forest: With five repetitions per context x_i, we use grouped cross-validation (group = x_i) so repetitions of the same x_i never span train and test folds to be sure we prevent any leakage. We limit model depth to 6 with 10 minimum nodes per leaf to control for overfitting.

Table 4: **Conditional-independence (belief sufficiency) tests: kNN conditional mutual information (CMI) for raw vs. isotonic-calibrated elicited beliefs.** For each dataset/model pair, we report kNN-based estimates of conditional mutual information $I(A;\theta \mid p)$ using (i) the raw elicited belief p and (ii) the isotonic-calibrated belief p_{iso}, each with 95% confidence intervals. Both correspond to the conditional-independence null hypothesis $H_0 : I(A;\theta \mid p) = 0$ (equivalently, $A \perp \theta \mid p$).

	CMI (Raw p)		**CMI (Isotonic p_{iso})**	
Dataset / Model	**CMI**	**95% CI**	**CMI**	**95% CI**
Heart–GPT-Min	0.1454	[0.1119, 0.1789]	0.3088	[0.2711, 0.3464]
Heart–GPT-High	0.0753	[0.0422, 0.1085]	0.3295	[0.2967, 0.3624]
Heart–Llama	0.0718	[0.0365, 0.1070]	0.1011	[0.0650, 0.1373]
Heart–DeepSeek	0.0675	[0.0354, 0.0997]	0.1948	[0.1604, 0.2291]
Cry–GPT-Min	0.2232	[0.1874, 0.2589]	0.2589	[0.2247, 0.2931]
Cry–GPT-High	0.1901	[0.1390, 0.2412]	0.2002	[0.1499, 0.2505]
Cry–Llama	0.4223	[0.3764, 0.4681]	0.5857	[0.5340, 0.6375]
Cry–DeepSeek	0.1745	[0.1265, 0.2225]	0.2110	[0.1673, 0.2547]
Fever–GPT-Min	0.1446	[0.1128, 0.1765]	0.1918	[0.1593, 0.2242]
Fever–GPT-High	0.0944	[0.0564, 0.1323]	0.1112	[0.0718, 0.1505]
Fever–Llama	0.3289	[0.2893, 0.3686]	0.5584	[0.5167, 0.6001]
Fever–DeepSeek	0.2060	[0.1663, 0.2456]	0.2407	[0.1947, 0.2867]
Diab–GPT-Min	0.0193	[0.0129, 0.0258]	0.0300	[0.0222, 0.0378]
Diab–GPT-High	0.0461	[0.0182, 0.0740]	0.0340	[0.0013, 0.0667]
Diab–Llama	0.2695	[0.2357, 0.3033]	0.4521	[0.4044, 0.4998]
Diab–DeepSeek	0.0351	[0.0169, 0.0533]	0.0353	[0.0162, 0.0544]

G Effect of Isotonic Calibration on Conditional Independence Tests

To evaluate whether probability calibration can mitigate violations of belief sufficiency, we repeated the conditional-independence analysis using isotonic regression to post-process elicited belief probabilities. Specifically, we replaced the raw elicited belief p with its isotonic-calibrated counterpart p_{iso} and re-estimated the conditional mutual information $I(A;\theta \mid p_{\text{iso}})$ using the same kNN-based estimator and bootstrap procedure.

Table 4 reports side-by-side comparisons of CMI estimates obtained from raw and isotonic-calibrated beliefs. Across all datasets and models, the qualitative conclusions remain unchanged: conditional mutual information remains substantially greater than zero, indicating persistent violations of the conditional-independence null hypothesis $H_0 : A \perp \theta \mid p$. In many cases, isotonic calibration increases the estimated CMI magnitude, reflecting improved marginal calibration without eliminating residual dependence between actions and ground truth after conditioning on elicited beliefs.

These results suggest that miscalibration alone does not account for the observed belief insufficiency. Instead, the dependence appears to arise from structural mismatches between elicited beliefs and the internal decision-relevant representations used by the models. Consequently, monotonic post-hoc calibration methods such as isotonic regression are insufficient to restore conditional independence in this setting.

H Robustness of the Belief-Sufficiency and Monotonicity Tests

This appendix collects additional analyses that test whether the conclusions of Section 5 depend on evidence-dependent preferences, on the specifics of the estimators, on the elicitation prompt, or on the three-action decision task. Throughout, "rejection" refers to the belief-sufficiency null $H_0 : A \perp \theta \mid p$ and $\widehat{\gamma}$ to the signed monotonicity margin of Table 2.

H.1 Preferences Cannot Account for the Violations

Proposition 1 quantifies over every utility $u(a, \theta)$ and every regularizer, so no preference over actions and outcomes, however cautious or asymmetric across diseases, can by itself produce a conditional-independence violation. A remaining concern is a utility that varies with the evidence x. Restricting u to (a, θ) is deliberate: were u allowed to depend freely on x jointly with θ, any action distribution could be rationalized without the agent holding beliefs about θ at all, and no test for belief validity could have content. We can nonetheless show empirically that reasonable x-dependent preferences are not what produces our result, in two ways (Table 5).

Stratification. We recompute the CI test within strata of each evidence covariate (stratifying on one covariate at a time), permuting actions only within stratum $\times$ belief bins, for the 147 covariate-stratifications with adequate support across the 16 dataset–model pairs. 110 of 147 stratifications reject (75%, against 5% expected under the null; mean within-stratum CMI 0.105), and 12 of 16 pairs reject under a majority of their stratifications; the only pair rejecting under none is Diabetes–DeepSeek, where the main analysis already reports the smallest violation. (Stratifying on all covariates at once is uninformative: the model never sees θ, so with the full context x held fixed its action is independent of θ by design and the test would pass for any model.)

Explicitly stated payoffs. We collected decisions under six stated payoff matrices $(c_{\mathrm{FP}}, c_{\mathrm{FN}}, c_{\mathrm{defer}})$, with the false-positive, false-negative, and deferral costs written into the decision prompt and beliefs elicited separately in a fresh context as in the main experiments. Here the utility is specified rather than inferred, so there is no room for ambiguity about the model's preferences. Violations persist: 13 of 15 testable pairs reject when a pair's six scenarios are pooled (conditioning on the stated cost configuration alongside the belief so that the utility stays fixed within every stratum), and 51 of 90 individual scenario cells reject on only 200 decisions each (Table 6). The excluded pair is Diabetes–GPT-Min, which defers on 99.5% of these decisions and so offers almost no action variation to test. The violations are smaller than in the main analysis (mean CMI 0.076 against 0.157), exactly as they should be: stating the costs turns the decision into a nearly mechanical threshold rule on a reported probability, which gives withheld information much less opportunity to affect the choice. Cyclic monotonicity continues to hold throughout ($\widehat{\gamma} \geq 0$ in 82 of 90 cells). As a check that the models attend to the stated numbers rather than ignoring them, raising the stated deferral cost from 0.5 to 2.0 lowers the mean deferral rate from 0.30 to 0.21 for GPT-Min and from 0.31 to 0.07 for GPT-High.

Table 5: **Belief sufficiency and monotonicity with preferences controlled.** Row 1 is the main analysis; rows 2–4 are the stratified and stated-payoff analyses and reach the same conclusion.

Analysis	**Testable cells**	**Sufficiency rejections**	**Mean CMI**	**Mean $\widehat{\gamma}$**	$\widehat{\gamma} \geq 0$
Main analysis (no stated costs)	16 pairs	16/16	0.157	0.075	15/16
Stratified by each supported covariate	147 tests	110/147	0.105	–	–
Explicit costs, per-scenario cells	90 cells	51/90	0.076	0.054	82/90
Explicit costs, pooled within pair	15 pairs	13/15	0.076	–	–

Table 6: **Both tests rerun with explicit costs stated in the decision prompt.** Beliefs are elicited separately; 15 dataset–model pairs are testable per scenario.

Stated $(c_{\mathrm{FP}}, c_{\mathrm{FN}}, c_{\mathrm{defer}})$	**Mean CMI**	**Rejections**	**Mean $\widehat{\gamma}$**	$\widehat{\gamma} \geq 0$
(1, 1, 0.5)	0.066	7/15	0.053	14/15
(1, 2, 0.5)	0.068	9/15	0.069	12/15
(2, 1, 0.5)	0.057	8/15	0.044	12/15
(1, 1, 2)	0.086	9/15	0.051	15/15
(1, 2, 2)	0.096	9/15	0.039	14/15
(2, 1, 2)	0.083	9/15	0.067	15/15
Overall (per-scenario cells)	0.076	51/90	0.054	82/90
Pooled within pair	0.076	13/15	–	–

H.2 The Finding Does Not Depend on the Estimator

We varied every component of the sufficiency test (Table 7). First, we vary the number of neighbours k used by the conditional-mutual-information estimator ($k = 3, 5, 10$). Second, we implemented a discrete plug-in alternative, which bins the variables and computes conditional mutual information from the resulting contingency table, replacing the kNN estimator entirely. Third, we replace the bootstrap criterion with a stricter permutation test: rather than resampling, we shuffle actions only among cases in the same decile of reported probability, which leaves the action–belief relationship intact and isolates whether actions still predict the outcome; this is the exact null corresponding to $A \perp \theta \mid p$. Under that test the three values of k reject 15, 14, and 14 of 16 pairs, and the plug-in estimator rejects 13 of 16. On repeated draws, we also re-ran the bootstrap resampling whole contexts rather than individual rows, so that all five repetitions of a context move together (the Random Forest comparison already grouped repetitions this way, Appendix F); 14 of 16 intervals remain above zero. Across every variant, each non-rejection is from the Diabetes domain, precisely where the paper already reports the smallest violations, and the ordering of dataset–model pairs is essentially unchanged: rank correlations between any two variants run from 0.86 to 0.96, so every estimator picks out the same models and domains as the large violators.

Monotonicity is likewise stable: the mean margin $\widehat{\gamma}$ stays positive for every bin count from 3 to 8, and a scan over bin counts from 2 to 10 finds at most 2 of 16 statistically significant violations at any bin count. On synthetic choice patterns with known violations, the statistic detects true violations with power above 95% at these sample sizes.

Table 7: **The sufficiency rejection survives changes to the estimator, the neighbour count, the null construction, and clustering by context** (16 dataset–model pairs). Rows 1, 2, and 6 share a point estimate because they use the same estimator and differ only in how significance is assessed. Absolute CMI levels depend on the estimator, while the ordering of pairs does not (see text).

Variant	Mean CMI	Rejections
Paper's criterion: $k = 3$, bootstrap CI > 0	0.157	16/16
$k = 3$, shuffle within belief bins	0.157	15/16
$k = 5$, shuffle within belief bins	0.118	14/16
$k = 10$, shuffle within belief bins	0.091	14/16
Discrete plug-in estimator, permutation test	0.181	13/16
Bootstrap resampling whole contexts ($k = 3$)	0.157	14/16

H.3 Alternative Elicitation Prompts Do Not Restore Sufficiency

We repeated both tests using beliefs from the MSE-scoring, absolute-loss, and Bayesian-reasoning prompts of Appendix B.1 (Table 8). No variant restores sufficiency, and monotonicity remains broadly satisfied under all of them. The Bayesian-reasoning instruction, which one might expect to help most, slightly *degrades* coherence. This supports the paper's message that the validity of an elicitation method has to be measured rather than assumed, and that the framework is what makes such measurement possible. All four rows are computed on the elicitation sample that contains the four prompt variants, using the $k = 3$ kNN estimator with the within-belief-bin permutation criterion (100 permutations) rather than the bootstrap interval of the main analysis; the standard-prompt row therefore differs slightly from the main-analysis row of Table 5.

Table 8: **Both tests under alternative elicitation prompts** (16 dataset–model pairs each; $k = 3$ kNN CMI with the within-belief-bin permutation criterion).

Elicitation prompt	Mean CMI	Rejections	Mean $\widehat{\gamma}$	$\widehat{\gamma} \geq 0$
Standard (π_0)	0.139	15/16	0.056	15/16
MSE scoring rule	0.161	15/16	0.064	14/16
Absolute loss	0.164	14/16	0.070	15/16
Bayesian reasoning	0.203	13/16	0.044	12/16

H.4 Deferral Responds to Reported Uncertainty

A defining feature of behavior that is driven by beliefs, as opposed to pure hedging, is that it responds systematically to those beliefs. Deferral decisions show such a pattern in three of the four models (Table 9). Ranking each pair's cases by how far the reported probability sits from 0.5, deferral is 117% more frequent in the most uncertain quarter of cases than in the most confident quarter for GPT-High, 51% more frequent for GPT-Min, and 32% for DeepSeek. Llama is the exception: its deferral runs *backwards* (-9%), so it defers less on the cases it reports as most uncertain. Llama is also the model our tests flag most often, which is the framework doing its job: the goal is not for every model's numbers to look good, but to distinguish models whose reported beliefs are plausibly valid from models whose reports are not, and Llama is the clearest case of the latter.

Table 9: **Deferral rises with reported uncertainty for every model but Llama.** Cases are split into quartiles by the distance of the reported probability from 0.5; the table gives the relative increase in deferral from the most confident to the most uncertain quartile, pooling decisions over the four domains.

Model	Deferral increase, confident $\rightarrow$ uncertain
GPT-Min	$+51\%$
GPT-High	$+117\%$
Llama	-9%
DeepSeek	$+32\%$

H.5 An Exploratory Extension with an Information-Gathering Action

Our main decision task has three actions. To probe whether the framework and its conclusions carry over to richer tasks, we built a minimal extension on the two public domains in which the model may also *order a test*: a focused echocardiogram measurement for Heart and a brief clinical assessment for Diabetes. A model that orders the test sees its result and then makes a final diagnostic decision, so the action space has four elements (diagnose No, diagnose Yes, defer, order a test). We ran GPT-5 at minimal and high reasoning over 100 contexts per domain, two repetitions each, in two regimes (with and without stated costs). Without stated costs, as in our standard prompt, the paper's conclusions carry over (Table 10): sufficiency is rejected in Heart for both models ($p < 0.005$ under the within-belief-bin permutation test), neither model rejects in Diabetes, the domain in which the paper already finds reported beliefs closest to sufficient, and $\widehat{\gamma} \geq 0$ everywhere. When costs are stated, test-ordering responds like a genuine action: as the stated test cost rises, the high-reasoning model's ordering share falls from 0.84 to 0.36. We view this as an expository test of the framework in an expanded setting rather than a full study; a systematic treatment of information-gathering actions, where the value of information depends on the belief itself, is a natural direction for future work.

Table 10: **Both tests on the information-gathering task in the regime without stated costs, GPT-5.** CMI is $I(A;\theta \mid p)$ over the four-action choice, with the within-belief-bin permutation p-value (200 permutations).

Reasoning	Domain	n	CMI	Perm. p	$\widehat{\gamma}$
Minimal	Heart	200	0.094	< 0.005	0.03
Minimal	Diabetes	200	-0.038	0.94	0.00
High	Heart	200	0.042	< 0.005	0.00
High	Diabetes	200	0.007	0.54	0.00

I COMPUTE

No GPUs were used. We accessed LLMs through API Keys, incurring roughly 10,000 in API fees. A standard CPU was able to process all data within minutes.