DO LLMS ACT LIKE RATIONAL AGENTS? MEASURING BELIEF COHERENCE IN PROBABILISTIC DECISION MAKING

Khurram Yamin[1*] **Jingjing Tang**[1] **Santiago Cortes-Gomez**[1]
Amit Sharma[2] **Eric Horvitz**[2] **Bryan Wilder**[1]
[1]Carnegie Mellon University [2]Microsoft Research

ABSTRACT

Large language models (LLMs) are increasingly deployed as agents in high-stakes domains where optimal actions depend on both uncertainty about the world and consideration of utilities of different outcomes, yet their decision logic remains difficult to interpret. We study whether LLMs are rational utility maximizers with coherent beliefs and stable preferences. We consider behaviors of models for diagnosis challenge problems. The results provide insights about the relationship of LLM inferences to ideal Bayesian utility maximization for elicited probabilities and observed actions. Our approach provides falsifiable conditions under which the reported probabilities *cannot* correspond to the true beliefs of any rational agent. We apply this methodology to multiple medical diagnostic domains with evaluations across several LLMs. We discuss implications of the results and directions forward for uses of LLMs in guiding high-stakes decisions.

1 INTRODUCTION

LLMs are increasingly used to support high-stakes decisions under uncertainty, such as medical diagnosis, where good actions require reasoning about a patient's latent state and trading off costs and benefits. To understand and improve these recommendations, a common strategy is to elicit the model's probabilities over relevant unknowns (e.g., disease likelihoods) from the available evidence. For instance, developers may wonder whether failures to seek additional information Liu et al. (2024b); Johri et al. (2025) stem from overconfident beliefs. However, it is unclear whether stated probabilities reflect the model's "true" beliefs: an elicited probability could track an internal epistemic state, or it could be a superficial linguistic output only weakly linked to the computations that drive choices (Pal et al., 2025; Wang et al., 2024a; Liu et al., 2024a). Making precise what it means for an LLM to hold a belief – formally, a subjective probability—and how to test it remains challenging.

Several recent studies, focus on testing the accuracy and calibration of expressed probabilities (Ulmer et al., 2024; Wang et al., 2024b; Cruz et al., 2024). While important, these questions are distinct. On the one hand, a decision maker can hold inaccurate or miscalibrated beliefs in good faith. On the other hand, LLMs could express beliefs that fail such conditions for many reasons, even if they internally "know better." For example, biases introduced during finetuning could lead models to verbalize miscalibrated probabilities.

We propose a novel strategy for validating elicited probability judgments from LLMs that centers on jointly eliciting *beliefs* and *decisions* on tasks where action quality depends on the relevant probabilities. E.g., in medical diagnosis, we ask models both for the probability that a patient has a condition and for a diagnostic action (e.g., yes/no/abstain). We then test whether the elicited probabilities could be the subjective beliefs of *any* rational agent for that decision problem. In the axiomatic tradition of Neumann & Morgenstern (1944), rational action under uncertainty corresponds to expected-utility maximization with respect to a belief over states and a utility function over outcomes. Building on Bayesian decision theory, we treat beliefs as subjective probabilities while remaining agnostic about utilities (McFadden, 1973), deriving testable implications that link reported beliefs to choices;

*Correspondence to: kyamin@andrew.cmu.edu

violations imply the joint pattern of elicited beliefs and actions cannot be rationalized by any Bayesian utility maximizer.

This yields a minimalist but falsifiable evaluation strategy: rather than asserting that a reported probability reflects the model's "true belief," we identify conditions under which it *cannot* be the belief of a rational decision-maker. Violation rates then quantify how well utility maximization (relative to elicited beliefs) explains behavior. In this sense, we propose a decision-oriented means to sidestep defining what a model "really" believes: when the goal is ultimately to understand and improve model *behavior*, elicited beliefs are a useful measure of internal state if they induce a self-consistent description of how the model makes decisions. Our empirical strategy is fully black-box, using only elicited probabilities and observed choices, and is complementary to white-box approaches that analyze internal activations. Finally, compared to prior work that tests probabilistic self-consistency (e.g., adherence to probability axioms), our approach yields richer, decision-linked implications.

Finally, we empirically instantiate this framework across multiple LLM families on clinically grounded diagnostic tasks spanning four medical settings. Beyond conventional calibration and discrimination, we evaluate whether elicited diagnostic beliefs are action-consistent, providing a sharper test of whether reported uncertainty is mechanistically meaningful for high-stakes decision support. This complements current efforts toward uncertainty-aware clinical LLMs that explicitly model diagnostic uncertainty and its explanation (Zhou et al., 2025).

2 Related Work

Belief elicitation from humans: A long literature in psychology and economics seeks to elicit true subjective probabilities, often using monetary incentives to encourage truthful reporting. Yet reviews Schlag et al. (2015); Charness et al. (2021) emphasize that validating belief elicitation is difficult and arguably ill-posed: beliefs are unobservable and may be shaped by elicitation itself, so researchers frequently evaluate methods by how well elicited beliefs predict actions. Our approach is similar in spirit, but LLMs pose different constraints. Human studies typically assume preferences (e.g., more money is 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 model in simulation is difficult. Accordingly, we 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 been 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 tell us in verbalized probabilities. Mechanistic interpretability uses access to the models' internal representations to test such questions Sharkey et al. (2025); Bereska & Gavves. For example, researchers train probes which predict distinctions like true vs false claims Azaria & Mitchell (2023); Marks & Tegmark (2023) from activations. We view mechanistic interpretability as complementary to our work in two respects: (1) our tests are fully black-box and rely only on model outputs (useful for closed models), and (2) 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. Our framework is unsupervised and output-only, and we view multiple toolkits as valuable given the difficulty of formally defining "belief."

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 approach such questions.

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, a decision maker forms its own (potentially misspecified) *subjective* posterior belief $P_S(\theta \mid x)$, which need not equal $P^\star(\theta \mid x)$. In the classic framework of expected utility maximization, the decision maker then selects an action $a \in \mathcal{A}$ to maximize expected utility:

$$a(x) \;=\; \arg\max_{a \in \mathcal{A}} \sum_{\theta \in \Theta} P_S(\theta|x) u(a, \theta). \tag{1}$$

Equivalently, letting $\ell(a, \theta) = -u(a, \theta)$ denote a loss function, we can minimize the objective. 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 (or, equivalently, the loss ℓ).

Our goal is to test whether an agent's reported beliefs are consistent with their acting as a utility maximizer in order to tell whether it is possible for the reported beliefs to play the role of true subjective probabilities in rational decision making. However, testing for consistency with expected utility maximization is by itself too strict – agents may fail to exactly maximize their utility function for a variety of reasons, or their utility on a given instance may include idiosyncratic preferences that we do not observe. Accordingly, we base our empirical strategy on the *random utility model* which forms the foundation for econometric work on human behavior McFadden (1973). Random utility maximizers augment Equation 1 with an additional random noise term which generates variation in their decision making:

Definition 1 (Random Utility Model)**.** *An agent is a random utility maximizer if it selects actions according to*

$$a(x) \in \arg\max_{a \in \mathcal{A}} \sum_{\theta \in \Theta} \left(P_S(\theta|x) u(a, \theta)\right) + \varepsilon_a, \tag{2}$$

where $\{\varepsilon_a\}_{a \in \mathcal{A}}$ *are action-specific random variables which satisfy* $\{\varepsilon_a\}_{a \in \mathcal{A}} \perp (x, \theta, P_S(\theta \mid x))$.

The randomness in ϵ induces (from the perspective of the experimenter) randomness in the choices of the agent. For example, if ϵ follows a logistic distribution (the most common specification in the literature), we obtain

$$\Pr(a(x) = a) \propto \exp\left(\sum_{\theta \in \Theta} P_S(\theta|x) u(a, \theta)\right). \tag{3}$$

That is, from the experimenter's perspective, the agent takes a softmax of their utility function. While this is one common distributional model that we include for intuition, our results will not require specific parametric assumptions on ϵ.

A final common modification to the standard utility-maximization framework is to allow agents to express preferences over risk that are not compatible with strict expected utility maximization. For example, humans are often observed to act in risk-averse ways, or to express greater sensitivity to losses than gains. Some work has argued that LLMs exhibit similar phenomena (Jia et al., 2024;

Payne, 2025; Hintze et al., 2025). Prospect theory Kahneman & Tversky (1979) is the classical modification of expected utility maximization to account for such behavior. In prospect theory, the agent has a monotone *probability weighting* function $w : [0, 1] \to [0, 1]$ and calculates losses for each action with respect to $w(P_S)$ instead of P_S, maximizing $\sum_{\theta\in\Theta} w(P_S(\theta|x))u(a, \theta)$. Combining this with random shocks to account for variation in decision making De Palma et al. (2008), we obtain

Definition 2 (PT-RUM)**.** *An agent is a prospect-theoretic random utility maximizer if it selects actions according to*

$$a(x) \in \arg\max_{a\in\mathcal{A}} \Big(\sum_{\theta\in\Theta} w(P_S(\theta|x))u(a, \theta) + \varepsilon_a \Big), \tag{4}$$

where w is monotone increasing and $\{\varepsilon_a\}_{a\in\mathcal{A}} \perp (x, \theta, P_S(\theta \mid x))$.

Our empirical strategy will allow for the possibility that agents act in a risk-averse manner consistent with Definition 2, in order to separate whether it is possible for elicited beliefs to represent subjective probabilities from whether the agent is risk-averse.

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 probability beliefs to reflect *subjective* belief states that actually drive its choices in the manner prescribed by rational decision making. 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 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 either knows the true joint distribution over (x, θ) or has a sample from this distribution.

Our framework gathers 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 the validity of. In our experiments, we prompt the model to report a probability distribution $P_E(\theta \mid x)$ in natural language (Paruchuri et al., 2024). Specifically, for each context x in our evaluation set, we query the model to provide probability estimates for each possible state. For the tasks that 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. The model is instructed to provide a numerical probability estimate. However, our methods are applicable to evaluate any elicitation method. We select natural-language probability estimates as our experimental focus because of their simplicity and the wide literature investigating them.

As a notational note, we will often use $p(x) := P_E(\theta = 1 \mid x)$ or $p^*(x) := P^*(\theta = 1 \mid x)$ in shorthand to refer to 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). Third, we apply tests examining properties that the sample $(x_i, P_E(\theta = 1 \mid x_i), a_i, \theta_i)$ should satisfy if the elicited beliefs and actions indeed reflect rational decision making relative to the expressed beliefs. Violations of these properties quantify the extent to which elicited beliefs depart from useful interpretation as a true subjective probability. We next detail these tests.

Conditional independence of actions and outcomes: The key intuition behind this test is that, once we condition on model's subjective belief, the realized outcome θ provides no additional information that could affect the choice: P_S is a sufficient statistic for the impact of the distribution of future events on the agent's choices. On the other hand, if a rational agent's actions have excess correlation with the true state given their elicited P_E, the agent must have information about θ that was not reported.

Proposition 1 (Conditional Independence under Truthful Reporting)**.** *For any agent satisfying Definition 1 or 2, 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 or 2 (under any utility function) who holds subjective belief $P_S = P_E$.*

For intuition, fix x. Under Definition 2, the expected loss depends on x only through $w(P_E(\theta \mid x))$, so Eq. equation 4 implies

$$a(x_i) = g(w(P_E(\theta \mid x)), \varepsilon)$$

for some function g. Since $\varepsilon \perp (x_i, \theta, P_E(\theta \mid x))$, we obtain $a \perp\!\!\!\perp \theta \mid w(P_E(\theta \mid x))$ and specifically $a \perp\!\!\!\perp \theta \mid P_E(\theta \mid x)$.

Remark 1. Observing $a \perp\!\!\!\perp \theta \mid P_E(\theta \mid x)$ does *not* imply 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 being a utility-maximizing agent with subjective belief $p_S = p_E$.

In order to empirically test this condition, we first conduct a direct conditional-independence (CI) test of the null hypothesis $A \perp\!\!\!\perp \theta \mid w(P_E(\theta \mid x))$ 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 agent's action a and compare nested feature sets. *Model 1* conditions on the elicited belief $p_E(x)$, while *Model 2* conditions on $(p_E(x), \theta)$. Under the conditional-independence null, adding θ should not improve prediction once we condition on p_E. We use cross-validated log-loss and report percent improvement as the relative reduction in log-loss from including θ. Significance is assessed by a bootstrap CI. To check whether violations are driven by omitted context, we then repeat the test while conditioning on an explicit evidence vector x: *Model 1^x* uses $(p_E(x), x)$ and *Model 2^x* uses $(p_E(x), x, \theta)$. Persistent gains from θ then indicate that $p_E(x)$ is not decision-sufficient even after controlling for observed context.

Monotone pairwise choice probabilities. In this section, we specialize to binary classification problems with state $\theta \in \{0, 1\}$ (as in all of the empirical examples we consider). The intuition behind this test is that, while we do not know the agent's utilities, the difference in their utilities for different actions is a monotone function of p. For example, an expected utility maximizer evaluates the difference in utility between two actions a_1 and a_2 as

$$\begin{aligned}\Delta_{a_1,a_2}(p_S(x)) = p_S(x) \cdot (u(a_1, 1) - u(a_2, 1)) \\ + (1 - p_S(x)) \cdot (u(a_1, 0) - u(a_2, 0))\end{aligned}$$

which is a linear function of $p_S(x)$ (and hence monotone). Monotonicity of pairwise differences in utility translates into monotonicity in choice probabilities under an additional restriction:

Definition 3 (IIA)**.** *Let* $\Pr(a(x) = a|\mathcal{C})$ *denote the probability that an agent chooses action* a *when presented with choices* $\mathcal{C} \subseteq \mathcal{A}$*. The agent satisfies independence of irrelevant alternatives if for any actions* $a_1, a_2 \in \mathcal{A}$ *and any* $\mathcal{C}$ *containing* a_1, a_2*,* $\frac{\Pr(a(x)=a_1|\mathcal{C})}{\Pr(a(x)=a_2|\mathcal{C})} = \frac{\Pr(a(x)=a_1|\mathcal{A})}{\Pr(a(x)=a_2|\mathcal{A})}$.

Essentially, IIA requires that the relative probability of the agent choosing a over b does not depend on the presence of other items. It is satisfied for example by the logit model in Equation 3. Note that it is also automatically satisfied for decision problems with two actions. Under this condition, we obtain a testable condition that the relative choice probabilities between two actions are monotone in $p_S(x)$:

Proposition 2 (Monotone pairwise choice probability)**.** *For any agent satisfying either Definition 1 or 2 in conjunction with IIA (Definition 3) the* pairwise (conditional) choice probability *for any two actions* $a_1, a_2 \in \mathcal{A}$*,*

$$s(p) \;:=\; \Pr\big(a(x) = a_1 \,\big|\, a \in \{a_1, a_2\},\, p_S(x)\big)$$

is either monotone increasing or monotone decreasing in the subjective probability p_S*.*

Remark 2. It is possible to test implications of random utility models for choice probabilities over more than two actions simultaneously via a generalized notion referred to as *cyclic monotonicity* McFadden (1981). This requires that there exist a convex function which has the choice probabilities as a subgradient. Since testing and interpreting this property is considerably more complex, we recommend pairwise comparisons here.

To test Proposition 2, we discretize elicited beliefs into K quantile bins $B_1, \ldots, B_K$ with non-decreasing centers $\bar{p}_1 \leq \cdots \leq \bar{p}_K$. For each bin B_k, we estimate the pairwise share of a_1 among

Table 1: **Conditional-independence (belief sufficiency) tests: kNN conditional mutual information (CMI) and CatBoost (CB) predictive-improvement tests (with and without explicit evidence features).** For each dataset/model pair, the first block reports the kNN-based estimate of conditional mutual information $I(A;\theta \mid p)$ (CMI) with 95% confidence intervals, corresponding to the conditional-independence null hypothesis $H_0 : I(A;\theta \mid p) = 0$ (equivalently, $A \perp \theta \mid p$). The second block reports CatBoost comparisons of $A \sim p$ vs. $A \sim (p,\theta)$, summarized by percent log-loss improvement with 95% bootstrap CIs. The third block reports the analogous CatBoost comparison additionally conditioning on the explicit evidence/context vector x ($A \sim (p,x)$ vs. $A \sim (p,x,\theta)$).

	kNN CMI: $I(A;\theta \mid p)$	**CB:**$A \sim p$ **v.** $A \sim (p,\theta)$	**CB:**$A \sim (p,x)$ **v.** $A \sim (p,x,\theta)$
Dataset / Model	**CMI 95% CI**	**% Impr 95% CI**	**% Impr 95% CI**
Heart–GPT-Min	0.1454 [0.1119, 0.1789]	16.37 [11.85, 20.90]	13.05 [9.16, 17.54]
Heart–GPT-High	0.0753 [0.0422, 0.1085]	4.23 [0.77, 7.62]	3.82 [1.07, 6.24]
Heart–Llama	0.0718 [0.0365, 0.1070]	2.98 [1.28, 5.07]	1.02 [-0.33, 2.39]
Heart–DeepSeek	0.0675 [0.0354, 0.0997]	4.52 [1.71, 7.48]	2.03 [0.02, 4.00]
Cry–GPT-Min	0.2232 [0.1874, 0.2589]	14.84 [12.08, 17.73]	4.88 [2.20, 7.55]
Cry–GPT-High	0.1901 [0.1390, 0.2412]	11.52 [8.80, 14.08]	9.25 [6.97, 11.34]
Cry–Llama	0.4223 [0.3764, 0.4681]	0.56 [-1.27, 2.48]	2.83 [1.17, 4.56]
Cry–DeepSeek	0.1745 [0.1265, 0.2225]	0.88 [-1.06, 2.54]	2.66 [1.04, 4.11]
Fever–GPT-Min	0.1446 [0.1128, 0.1765]	12.67 [10.06, 15.53]	9.45 [7.14, 11.71]
Fever–GPT-High	0.0944 [0.0564, 0.1323]	4.74 [2.11, 7.13]	2.27 [0.25, 4.19]
Fever–Llama	0.3289 [0.2893, 0.3686]	8.38 [6.11, 10.78]	4.75 [2.40, 7.21]
Fever–DeepSeek	0.2060 [0.1663, 0.2456]	22.83 [19.85, 26.06]	15.83 [12.85, 18.81]
Diab–GPT-Min	0.0193 [0.0129, 0.0258]	2.84 [0.07, 5.94]	0.00 [-0.01, 0.01]
Diab–GPT-High	0.0461 [0.0182, 0.0740]	0.22 [-0.36, 0.75]	0.43 [-0.77, 1.69]
Diab–Llama	0.2695 [0.2357, 0.3033]	12.33 [9.12, 15.39]	-0.26 [-0.73, 0.12]
Diab–DeepSeek	0.0351 [0.0169, 0.0533]	-2.43 [-3.01, -1.90]	0.01 [0.01, 0.02]

$\{a_1, a_2\}$:

$$\widehat{s}_k := \frac{\#\{i \in B_k : A_i = a_1\}}{\#\{i \in B_k : A_i \in \{a_1, a_2\}\}}. \tag{5}$$

We flag a monotonicity violation for any pair of bins $j < k$ where $\widehat{s}_j > \widehat{s}_k$. For each flagged pair, we assess significance using a one-sided test for proportions (e.g., Fisher's exact test or a one-sided binomial test) of $H_0 : s(\bar{p}_j) \leq s(\bar{p}_k)$, and report the fraction of pairwise comparisons with significant violations at $\alpha = 0.05$.

Consistency across decision tasks. For a rational agent, beliefs and preferences are distinct: when multiple decision tasks reference the same state θ, agents adhering to Definitions 1 or 2 should act as if driven by a common subjective belief p_S even if the action set and utility differ across tasks. 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, beyond our single-task tests, we compare elicited probabilities p_E across prompts that embed different decision tasks but share the same (x, θ). If the model truthfully reports a stable belief, these elicited probabilities should agree. In binary settings, we instantiate alternative tasks by asking for $p^*(\theta \mid x)$ under different stated evaluation losses (e.g., MSE or MAE).

We test consistency via repeated elicitations. Let $p(x;\pi,j)$ denote the reported $P_E(\theta \mid x)$ for context x under prompt for task $\pi \in \Pi$ on repetition j, and let $\bar{p}(x;\pi) = \frac{1}{r}\sum_{j=1}^{r} p(x;\pi,j)$ denote the repetition-averaged elicited probability. Fixing the main task π_0 (e.g., the medical diagnosis tasks we construct), for each alternative $\pi \in \Pi \setminus \{\pi_0\}$ (with Π the set of prompts), we compute the RMSE between $p(x;\pi)$ and $\bar{p}(x;\pi_0)$. Larger RMSE indicates less consistency, though as with any of our tests we do not expect perfect consistency. For example, Ronayne et al. (2022) tested prize independence in humans and found that over half of subjects failed. For reference, for the standard prompt π_0, we also show the standard deviation across repetitions.

Internal Consistency via Law of Iterated Expectation. To validate whether elicited beliefs constitute a coherent probability distribution, we test adherence to the law of iterated expectation over an auxiliary variable z (e.g., a patient attribute) that we can optionally reveal. The tests in

Table 2: **Monotone pairwise choice: significant violation rates.** Fraction of bin-pair comparisons with a statistically significant (one-sided test at $\alpha = 0.05$) monotonicity violations for three action pairs: (Yes, No), (Yes, Defer), and (Defer, No) with Yes ($a = 1$) being represented as Y, No($a = 0$) as N, and Defer as D. All values are reported in percentage units.

Dataset/Model	Y/(N+Y)	Y/(D+Y)	D/(N+D)
Heart–GPT-Min	0.0	0.0	0.0
Heart–GPT-High	0.0	0.0	0.0
Heart–Llama	0.0	0.0	0.0
Heart–DeepSeek	0.0	20.0	30.0
Cry–GPT-Min	0.0	0.0	10.0
Cry–GPT-High	0.0	0.0	0.0
Cry–Llama	0.0	0.0	10.0
Cry–DeepSeek	0.0	0.0	0.0

Dataset/Model	Y/(N+Y)	Y/(D+Y)	D/(N+D)
Fever–GPT-Min	0.0	0.0	30.0
Fever–GPT-High	0.0	0.0	0.0
Fever–Llama	0.0	10.0	0.0
Fever–DeepSeek	0.0	0.0	30.0
Diab–GPT-Min	0.0	0.0	0.0
Diab–GPT-High	0.0	0.0	0.0
Diab–Llama	50.0	33.3	0.0
Diab–DeepSeek	0.0	10.0	0.0

earlier sections relate to coherence between beliefs and decisions. Here, we test a more purely *probabilistic* notion of coherence: a decision maker can "truthfully" report its subjective probabilities and have inconsistency between those beliefs. This check is a generalization of tests in prior work on axiomatic/self-consistency of LLM probability judgments (Zhu & Griffiths, 2025; Herrmann & Levinstein, 2024). We include it as an additional point of comparison to understand how probabilistic inconsistency relates to the decision-oriented violations above. Arguably, it may measure a different quantity: an agent could act according to Definition 1 even if their p_S does not add up to a valid distribution.

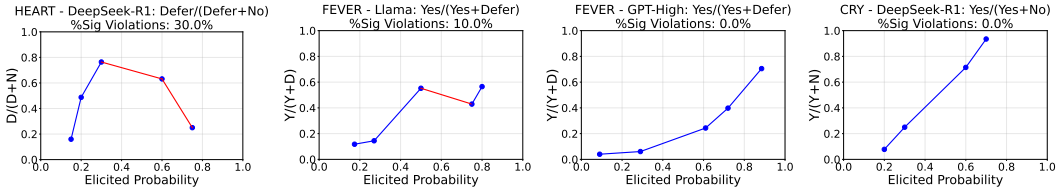
Concretely, let x denote the base context (with z withheld), and let $\{B_1, \ldots, B_k\}$ be a partition of the support of z. We verify: $P_E(\theta \mid x) = \sum_{j=1}^{k} P_E(\theta \mid x, z \in B_j)\ P_E(z \in B_j \mid x)$. In our experiments, we construct partitions based on patient covariates 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.

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 with five random seeds used to aggregate data from. 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.

Real-World 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 (Kahn, 1994). Ground-truth $p^*(x)$ are computed as the fraction of positive diagnoses within falling into covariate strata (always at least 100 patients).

Expert-Constructed Bayesian Networks. We also evaluate on distributions from two Bayes nets constructed by a leading expert in pediatric medicine. Due to data sharing restrictions, we do not release the full networks. However, pending completion of an data-sharing agreement, we plan to have these permissions by the camera-ready. We 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, lethargy). The crying network covers causes such as colic and gas pain with behavioral/physical covariates (e.g., feeding difficulties, distended abdomen).



(a) Heart-Deepseek (D,N) (b) Fever-Llama (Y,D) (c) Fever-GPT High(Y,D) (d) Cry-Deepseek (Y,N)

Figure 1: Monotone Odds Violations: In these plots, we show examples of cases with and without statistically significant monotonicity violations. The segments without violations are in blue and those with are in red. The x-axis shows the elicited probability, and the y-axis shows choice probability ratio. We use Y to represent Yes $(a = 1)$, N for No $(a = 0)$, and D for Defer.

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.

Prompting. 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 (e.g., explicit scoring rules or directions to perform Bayesian reasoning). Our emphasis is on the evaluation methodology rather than the exact wording, and prompts can be adapted.

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 realized outcome θ should add no predictive signal for A once we condition on the elicited belief (or an effective belief after probability weighting), i.e., $H_0 : I(A; \theta \mid p) = 0$. 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 5 shows the same conclusion for isotonic-calibrated beliefs, indicating calibration does not restore decision-sufficiency.

To gauge the size of this dependence, the right side of Table 1 reports two cross-validated CatBoost comparisons (details in Appendix) (Dorogush et al., 2018). Adding θ to $A \sim p$ yields out-of-sample log-loss improvements ranging from negligible ($< 1\%$) to moderate (10–20%), suggesting that p captures most—but not all—decision-relevant signal in many settings. Conditioning additionally on evidence x, improvements from adding θ to $A \sim (p, x)$ are smaller yet remain significant in most cases (12/16), implying residual belief–decision inconsistency not explained by observed context. Overall, elicited probabilities are often close (though not exact) sufficient statistics for decisions, motivating task- and model-specific validation.

Monotone Choice Probabilities. We apply the monotone probability test described in Section to three action pairs: (Yes, No), $(\text{Yes}, \text{Defer})$, and $(\text{Defer}, \text{No})$ with Yes being $(a = 1)$ and No being $(a = 0)$. By Proposition 2 and its assumptions, increasing the elicited belief in the positive diagnosis should have a monotonic non-decreasing effect on the odds of choosing the first item in each action pair. In Table 2, we show the fraction of pairwise bin comparisons with a statistically significant (one-sided test at $\alpha = 0.05$) monotonicity violations. To illustrate, Figure 1 shows examples that are test does and does not count as violations of monotonicity. Models consistently

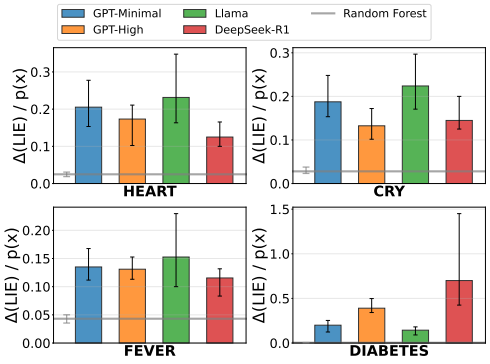


Figure 2: Internal Consistency Error via Law of Iterated Expectation (LIE): Here we plot the 95% CI for the median quantity $\Delta_{\text{LIE}}(x)/p(x)$ (see Section 3.2) by dataset(subplots) and model(bars). A cross-validated random forest model that is separately trained on ground truth $p^*(x)$ and $p^*(x, z)$ (also conditional on next state distribution z) values is used as a baseline.

have monotone pairwise choice probabilities for Yes/No choices, while some (model,dataset,bin) pairs exhibit violations for comparisons involving deferral (perhaps as the "in between" action). There is substantial heterogeneity across models, e.g., GPT-High always satisfies monotonicity.

Prompt Consistency. In Table 3 (Appendix), we measure standard deviation of the elicited beliefs under standard prompting, and measure RMSE deviations in elicited beliefs to various alternative prompts and decision tasks. These alternative prompts involve the LLM being told the scoring rule will be MSE, being told the scoring rule will be absolute loss, or being told to do Bayesian reasoning. We find that for most models (GPT-Min, GPT-High, DeepSeek) consistency across different decision tasks (MSE/MAE loss) is comparable to consistency across repeated samples for the same prompt. For Llama, changes in task description result in large differences in elicited probabilities. All models change behavior more substantially under the Bayesian reasoning prompt, consistent with that models may express beliefs more stably across decision tasks than across changes in "persona". GPT-High displays significantly less variability across all prompt variations than other models, while Llama exhibits the most by a considerable margin.

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) beliefs needed to form the LIE decomposition over an auxiliary "next-state" variable z. We assess probabilistic coherence via the Law of Iterated Expectation error $\Delta_{\text{LIE}}(x)$, summarized in Figure 2 by the median normalized ratio $\Delta_{\text{LIE}}(x)/p(x)$ (median for robustness when $p(x)$ is near 0). As a baseline, we compute the analogous quantity using cross-validated random forest predictors trained on ground-truth $p^*(x)$ and $p^*(x, z)$, yielding $\hat{\Delta}_{\text{LIE}}(x)/\hat{p}(x)$. Across datasets, LLMs typically exhibit substantially larger inconsistency than the baseline, with no model consistently dominating the others. This indicates that consistency with probability axioms may indeed be a different property than decision-consistency; models that have consistently higher levels of decision-consistency in the earlier three tests do not necessarily fare better in this measure.

6 Discussion and Conclusions

We introduce a decision-theoretic framework for testing whether elicited LLM probabilities function as subjective beliefs that drive actions. Using the formalism of utility maximization under uncertainty, we derive falsifiable constraints linking beliefs to choices and propose practical tests for detecting when verbalized probabilities cannot rationalize behavior for *any* Bayesian utility maximizer. Applying our methods to medical diagnosis across several domains and LLMs, we find that elicited beliefs and actions are provably not fully consistent with rational decision making. However, for many tasks and models the degree of deviation is mild, indicating that elicited beliefs could be a useful lens into behavior with appropriate validation. *Since our tests are falsification-oriented, we cannot prove that elicited probabilities really represent true beliefs*. An important future topic is further characterizing deviations from rationality to inform model development. Broadly, our work shows that applying the formal constraints of rational decision making can offer useful insight into model behavior.

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.

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.

Andre De Palma, Moshe Ben-Akiva, David Brownstone, Charles Holt, Thierry Magnac, Daniel McFadden, Peter Moffatt, Nathalie Picard, Kenneth Train, Peter Wakker, et al. Risk, uncertainty and discrete choice models. *Marketing Letters*, 19(3):269–285, 2008.

Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. Catboost: gradient boosting with categorical features support, 2018. URL https://arxiv.org/abs/1810.11363.

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/.

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

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.

Arend Hintze, Charu Bisht, Jory Schossau, and Ralph Hertwig. Nonlinear transformation of probabilities by large language models. *Computers in Human Behavior: Artificial Humans*, 6: 100227, 2025. ISSN 2949-8821. doi: https://doi.org/10.1016/j.chbah.2025.100227. URL https://www.sciencedirect.com/science/article/pii/S2949882125001112.

Jingru Jia, Zehua Yuan, Junhao Pan, Paul E. McNamara, and Deming Chen. Decision-making behavior evaluation framework for llms under uncertain context, 2024. URL https://arxiv.org/abs/2406.05972.

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.

Michael Kahn. Diabetes. UCI Machine Learning Repository, 1994. DOI: https://doi.org/10.24432/C5T59G.

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.

Ollie Liu, Deqing Fu, Dani Yogatama, and Willie Neiswanger. Dellma: Decision making under uncertainty with large language models, 2024a. URL https://arxiv.org/abs/2402.02392.

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*, 2024b.

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. Econometric models of probabilistic choice. *Structural analysis of discrete data with econometric applications*, 198272, 1981.

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

John Von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ, USA, 1944.

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

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.

Akshay Paruchuri, Jake Garrison, Shun Liao, John Hernandez, Jacob Sunshine, Tim Althoff, Xin Liu, and Daniel McDuff. What are the odds? language models are capable of probabilistic reasoning, 2024. URL https://arxiv.org/abs/2406.12830.

Kenneth Payne. An analysis of ai decision under risk: Prospect theory emerges in large language models, 2025. URL https://arxiv.org/abs/2508.00902.

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

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 Pe-

terson, 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.

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.

Cheng Wang, Gyuri Szarvas, Georges Balazs, Pavel Danchenko, and Patrick Ernst. Calibrating verbalized probabilities for large language models, 2024a. URL https://arxiv.org/abs/2410.06707.

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

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.

Shuang Zhou, Jiashuo Wang, Zidu Xu, Song Wang, David Brauer, Lindsay Welton, Jacob Cogan, Yuen-Hei Chung, Lei Tian, Zaifu Zhan, Yu Hou, Mingquan Lin, Genevieve B. Melton, and Rui Zhang. Uncertainty-aware large language models for explainable disease diagnosis, 2025. URL https://arxiv.org/abs/2505.03467.

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 (a random element in $\Delta(\Theta)$). Under either Definition 1 (RUM) or Definition 2 (PT–RUM), the agent selects an action according to

$$A \in \arg\max_{a \in \mathcal{A}} \big\{ V_a(B) + \varepsilon_a \big\},$$

where, in the RUM case,

$$V_a(B) \;=\; \sum_{\theta \in \Theta} B(\theta)\, u(a, \theta),$$

and in the PT–RUM case,

$$V_a(B) \;=\; \sum_{\theta \in \Theta} w(B(\theta))\, u(a, \theta),$$

for a monotone $w : [0,1] \to [0,1]$. Fix an arbitrary tie-breaking rule (e.g. choose the smallest-index maximizer). Then there exists a measurable function g such that

$$A \;=\; g(B, \varepsilon), \qquad \text{where } \varepsilon := (\varepsilon_a)_{a \in \mathcal{A}}.$$

By the defining assumption of both models,

$$\varepsilon \;\perp\!\!\!\perp\; (x, \theta, B),$$

and in particular $\varepsilon \perp\!\!\!\perp (\theta, B)$.

We now show $A \perp\!\!\!\perp \theta \mid B$. Let $S \subseteq \mathcal{A}$ be any measurable set. Then

$$\Pr(A \in S \mid B, \theta) \;=\; \Pr(g(B, \varepsilon) \in S \mid B, \theta).$$

Condition on $(B = b, \theta = t)$. Since ε is independent of (B, θ), the conditional distribution of ε given $(B = b, \theta = t)$ equals the conditional distribution of ε given $B = b$ (indeed it equals the unconditional distribution). Hence

$$\Pr(g(B, \varepsilon) \in S \mid B, \theta) \;=\; \Pr(g(B, \varepsilon) \in S \mid B),$$

so

$$\Pr(A \in S \mid B, \theta) = \Pr(A \in S \mid B) \quad \text{a.s.}$$

Because this holds for all measurable S, it follows that

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

i.e. $A \perp\!\!\!\perp \theta \mid P_S(\theta \mid x)$.

For the converse, argue by contrapositive. Suppose there existed an agent satisfying Definition 1 or Definition 2 and truthful reporting $P_S(\theta \mid x) = P_E(\theta \mid x)$ almost surely. Then with $B_E := P_E(\theta \mid x)$ we have $B_E = B$ almost surely, and the first part yields

$$A \perp\!\!\!\perp \theta \mid B_E \;=\; A \perp\!\!\!\perp \theta \mid P_E(\theta \mid x).$$

Therefore, if empirically $A \not\perp\!\!\!\perp \theta \mid P_E(\theta \mid x)$, no such agent can satisfy $P_S(\theta \mid x) = P_E(\theta \mid x)$ almost surely. □

A.2 Proof of Proposition 2

. Fix two actions $a_1, a_2 \in \mathcal{A}$ and write $p := p_S(x) \in [0,1]$ for the agent's subjective probability of $\theta = 1$. For each action $a \in \mathcal{A}$, define the (deterministic) belief-indexed value

$$V_a(p) := \sum_{\theta \in \{0,1\}} \pi(\theta; p)\, u(a, \theta), \qquad \text{where } \pi(1;p) = p,\ \pi(0;p) = 1-p \tag{6}$$

under Definition 1, and under Definition 2 define instead

$$V_a(p) := \sum_{\theta \in \{0,1\}} \tilde{\pi}(\theta; p)\, u(a, \theta), \qquad \text{where } \tilde{\pi}(1;p) = w(p),\ \tilde{\pi}(0;p) = 1-w(p). \tag{7}$$

In either case, the random utility representation in Definitions 1–2 implies that, when the feasible set of actions is $\mathcal{C} \subseteq \mathcal{A}$, the agent chooses

$$a(x) \in \arg\max_{a \in \mathcal{C}} \big\{ V_a(p) + \varepsilon_a \big\},$$

where $\{\varepsilon_a\}_{a \in \mathcal{A}} \perp (x, \theta, p_S(x))$, hence the joint distribution of $(\varepsilon_{a_1}, \varepsilon_{a_2})$ does not depend on p.

Step 1: Reduce the conditional share to a binary-choice probability (IIA). By Definition 3, for any choice set $\mathcal{C}$ that contains a_1, a_2 we have

$$\frac{\Pr(a(x) = a_1 \mid \mathcal{C}, p)}{\Pr(a(x) = a_2 \mid \mathcal{C}, p)} = \frac{\Pr(a(x) = a_1 \mid \mathcal{A}, p)}{\Pr(a(x) = a_2 \mid \mathcal{A}, p)}.$$

In particular, taking $\mathcal{C} = \{a_1, a_2\}$ and rearranging yields

$$\Pr\big(a(x) = a_1 \,\big|\, a \in \{a_1, a_2\},\, p\big) = \Pr\big(a(x) = a_1 \,\big|\, \{a_1, a_2\},\, p\big). \tag{8}$$

Thus it suffices to study the binary menu $\{a_1, a_2\}$.

Step 2: Express the binary-choice probability as a CDF evaluated at a utility difference. On the binary menu $\{a_1, a_2\}$, the event that the agent chooses a_1 is

$$\{V_{a_1}(p) + \varepsilon_{a_1} \geq V_{a_2}(p) + \varepsilon_{a_2}\} \iff \{\varepsilon_{a_2} - \varepsilon_{a_1} \leq \Delta(p)\},$$

where

$$\Delta(p) := V_{a_1}(p) - V_{a_2}(p). \tag{9}$$

Let $F(t) := \Pr(\varepsilon_{a_2} - \varepsilon_{a_1} \leq t)$ be the (weakly) increasing CDF of the shock difference. Since $(\varepsilon_{a_1}, \varepsilon_{a_2})$ is independent of p, the function F does not depend on p, and

$$\Pr\big(a(x) = a_1 \,\big|\, \{a_1, a_2\},\, p\big) = F(\Delta(p)). \tag{10}$$

Combining equation 8 and equation 10 gives

$$s(p) = F(\Delta(p)).$$

Step 3: Show $\Delta(p)$ is monotone in p. Write $\Delta u(\theta) := u(a_1, \theta) - u(a_2, \theta)$. Under Definition 1, using equation 6–equation 9,

$$\Delta(p) = p\,\Delta u(1) + (1-p)\,\Delta u(0) = \Delta u(0) + p\big(\Delta u(1) - \Delta u(0)\big), \tag{11}$$

which is an affine function of p and therefore is either (weakly) increasing, (weakly) decreasing, or constant on $[0,1]$.

Under Definition 2, the same calculation with p replaced by $w(p)$ yields

$$\Delta(p) = w(p)\,\Delta u(1) + (1-w(p))\,\Delta u(0) = \Delta u(0) + w(p)\big(\Delta u(1) - \Delta u(0)\big). \tag{12}$$

Because $w(\cdot)$ is monotone increasing, equation 12 implies that $\Delta(p)$ is (weakly) increasing in p if $\Delta u(1) - \Delta u(0) \geq 0$, (weakly) decreasing if $\Delta u(1) - \Delta u(0) \leq 0$, and constant if $\Delta u(1) = \Delta u(0)$.

Step 4: Conclude monotonicity of $s(p)$. Since F is a CDF, it is (weakly) increasing. Therefore $s(p) = F(\Delta(p))$ is the composition of a (weakly) increasing map F with a function $\Delta(\cdot)$ that is either (weakly) increasing, (weakly) decreasing, or constant. Hence $s(\cdot)$ is either (weakly) monotone increasing, (weakly) monotone decreasing, or constant on $[0,1]$. This proves the claim. □

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 repetition-averaged elicited probabilities under alternative prompts (being told in prompt that MSE/Absolute-loss scoring rule will be used or Bayesian-reasoning instruction) vs π_0

	π_0	**RMSE: diff π vs π_0**		
Dataset/Model	**Std**	**MSE**	**Abs**	**Bayes**
Heart–GPT-Min	.0786	.0915	.0797	.1268
Heart–GPT-High	.0591	.0456	.0436	.0682
Heart–Llama	.1592	.2403	.2415	.2697
Heart–DeepSeek	.0781	.0829	.0753	.2523
Cry–GPT-Min	.1696	.1454	.1581	.1929
Cry–GPT-High	.0968	.0808	.0737	.1655
Cry–Llama	.0986	.2090	.1988	.2437
Cry–DeepSeek	.0948	.0885	.0999	.2047
Fever–GPT-Min	.1469	.1242	.1179	.1445
Fever–GPT-High	.1003	.0675	.0610	.1432
Fever–Llama	.0880	.1716	.1703	.2339
Fever–DeepSeek	.0890	.0929	.0972	.1870
Diab–GPT-Min	.0451	.0724	.0715	.0865
Diab–GPT-High	.0996	.0860	.0751	.1854
Diab–Llama	.0404	.2520	.2581	.3176
Diab–DeepSeek	.1004	.1137	.1095	.1670
Mean: GPT-Min	**.1100**	**.1084**	**.1068**	**.1377**
Mean: GPT-High	**.0889**	**.0700**	**.0634**	**.1406**
Mean: Llama	**.0965**	**.2182**	**.2172**	**.2662**
Mean: DeepSeek	**.0905**	**.0945**	**.0955**	**.2028**
Overall mean	**.0965**	**.1228**	**.1207**	**.1868**

Table 4: **RMSE vs. ground truth by model.** Root mean squared error between elicited probabilities and ground-truth posteriors, averaged across datasets, for the standard prompt (π_0) and alternative elicitation prompts (being told in prompt that MSE/Absolute-loss scoring rule will be used or Bayesian-reasoning instruction)

Model	**Standard**	**MSE**	**Bayesian**	**Absolute**
GPT-Minimal	.2236	.2417	.2369	.2443
GPT-High	.2273	.2138	.2780	.2158
Llama	.3434	.3391	.2784	.3411
DeepSeek-R1	.2187	.2224	.2732	.2180

C Prompting Analysis

C.1 Prompt Consistency

In Table 3, we measure standard deviation of the elicited beliefs under standard prompting, and measure RMSE deviations in elicited beliefs to various alternative prompts and decision tasks. These alternative prompts involve the LLM being told the scoring rule will be MSE, being told the scoring rule will be absolute loss, or being told to do Bayesian reasoning.

C.2 Prompting Techniques vs Ground Truth

We were also curious whether the high deviation for the Bayesian prompt meant the model was somehow doing stronger reasoning and bringing hidden knowledge by being a good Bayesian Reasoner. However, we can see from Table 4 (Appendix) showing RMSE vs ground truth by prompting technique that this does not seem to be the case with Bayesian Reasoning elicited probabilities having an average RMSE versus ground truth of .26, higher than all other prompting variations that all have an average RMSE of .23. Further, it is interesting that all prompt variations are closer to each other than they are to the ground truth.

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 (Kahn, 1994). 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 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 PARAMATER DETAILS

For CatBoost: 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, and limit iterations to 1000 to control for potential overfitting.

Table 5: **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 5 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.