Reasoning under Uncertainty: Trustworthy LLMs for High Stakes Medicine

Yusuf Kesmen, LiGHT, EPFL

***Abstract*—Conversational clinical decision support is sequential decision-making under uncertainty, yet large language models conflate fluent generation with probabilistic reasoning: they are poorly calibrated and their inference is neither explicit nor auditable. Classical decision-support systems had these virtues but relied on hand-crafted knowledge. This thesis seeks clinical reasoners that are explicit, calibrated, and auditable without sacrificing the generality of learned models. We pursue three connected directions: separating reasoning from knowledge by pairing language models with explicit probabilistic inference; instilling decision-theoretic reasoning into the models themselves; and benchmarking reasoning against clinician decision pathways. In preliminary work, a modular system that uses the language model only as a language interface to an explicit, calibrated diagnostic engine already surpasses far larger autonomous models at a fraction of the cost.**

***Index Terms*—Clinical decision support, medical AI, large language models, LLM reasoning, reasoning under uncertainty, probabilistic inference, trustworthy AI.**

I. Introduction

THE prospect of deploying autonomous AI systems in high-stakes domains rests on a single demanding property: trustworthiness. Medicine is a paradigmatic such domain, and conversational clinical decision support (CCDS), in which a system converses with a clinician or patient, gathers evidence over multiple turns, and recommends a course of action, is among its most exacting instances. Properly framed, CCDS is sequential decision-making under uncertainty: the system must entertain competing hypotheses about an unobserved cause, gather the evidence that most reduces its uncertainty, commit only when the balance of risks warrants it, and recognise when it should instead defer to a human.

A trustworthy system for this task must do more than emit plausible recommendations. Because the task is reasoning under uncertainty, its requirements follow directly: the system should maintain an explicit probabilistic state over hypotheses, gather information in a principled, value-driven manner, and report calibrated confidence, knowing when to abstain or defer, while exposing auditable and verifiable reasoning chains. Crucially, in medicine in particular, it is not enough for a decision to be correct: one must be able to understand why it was reached, since this is what ultimately grounds clinical trust and accountability.

Large language models (LLMs) are appealing candidates for this role. Trained on vast corpora, they encode a remarkable breadth of medical knowledge and perform strongly on medical question-answering benchmarks [1], [2], [3], [4], [5]. Yet they are ill-suited to the probabilistic core of the task, in several related ways. First, they are inaccurate and overconfident when asked to produce numerical confidences and probabilities [6]. Second, they give no out-of-the-box interpretable or controllable account of how those estimates are derived [7]. Third, the estimates they do produce are coarse-grained, assigning near-identical probabilities to materially different conditions and so failing to resolve the close calls a diagnosis turns on [8]. And fourth, asked not merely to state such probabilities but to act on them, they depart from expected-utility-rational choice, making inconsistent and suboptimal decisions under uncertainty [9]. Together these shortcomings make it difficult to trust LLMs as components of large-scale, automatic decision-making, precisely the setting in which their breadth would be most valuable. The gap is sharpest where the task most resembles clinical practice: as interaction moves from single-turn questions to multi-turn dialogue, noisy exchanges, and multi-step information gathering, performance degrades, models hallucinate, and they lose the thread of the conversation over its course [10], [11]. A growing body of agentic and reasoning-oriented methods has sought to make LLMs reason more deliberately about the world, but the best systems still struggle with the sustained, sequential reasoning that high-stakes settings demand.

It is worth recalling that an earlier paradigm handled parts of this problem well. Classical clinical decision support systems (CDSS), the rule-based and probabilistic models, produced reliable, accurate, and inspectable clinical decisions [12], [13], [14]. Their weakness was not rigor but reach: their knowledge bases were hand-crafted, brittle, and costly to maintain, they did not generalize, and they offered no natural interface to unstructured clinical language. This contrast exposes a useful decomposition. Reasoning comprises two ingredients, knowledge and inference: a clinician must both know the diseases, their presentations, and their likelihoods (knowledge) and search among them for the best explanation of the observed evidence (inference). Classical CDSS keep these ingredients separate, with knowledge explicitly encoded and a distinct engine performing inference; the inference is an algorithm we can inspect and therefore trust, and the knowledge comes from human experts we can trust so long as we are willing to curate it, which together make these systems auditable but also rigid. Autoregressive LLMs instead fuse the two into a single optimization, which lends them generality but also hides and miscalibrates their inference. Were the full distribution of how expert clinicians reason and decide available to train on, this fusion might be harmless, since a model could simply absorb that reasoning end-to-end; but no such data exists, and in its absence the single optimization inherits precisely the opacity and miscalibration the classical separation was designed to prevent.

Viewed through this lens, clinical reasoning departs from the

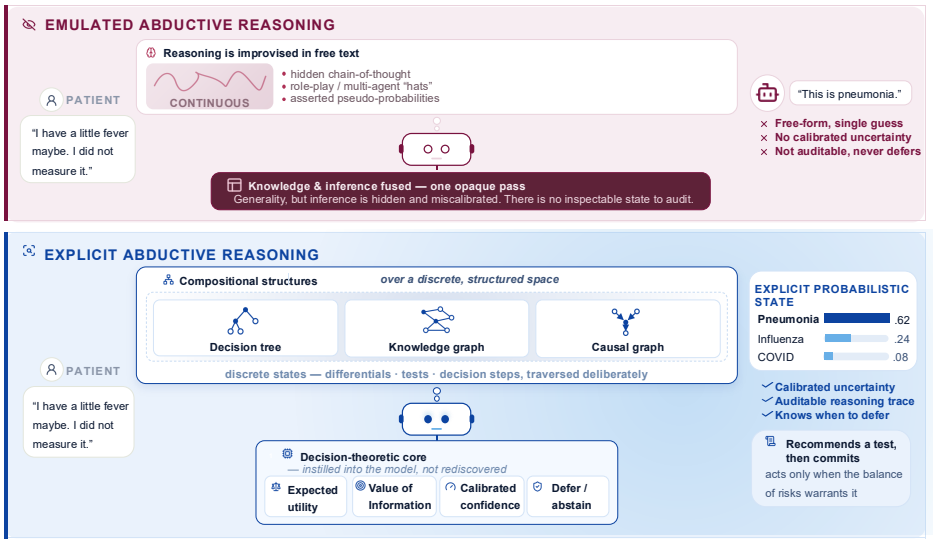

Fig. 1. Emulated versus explicit abductive reasoning. Today's language models fuse knowledge and inference into a single opaque pass, so their reasoning is implicit, uncalibrated, and unauditable; our approach instead equips the model with the structures clinicians actually reason with, drawn from data rather than hand-crafted, making inference explicit, structured, calibrated, and auditable.

dominant deep-learning paradigm in three respects, as Figure 1 illustrates. First, humans, clinicians included, are not naturally reliable reasoners under uncertainty; it is through cultural and scientific progress that we arrived at explicit normative theories for it, statistical decision and utility theory [15], [16], [17], developed precisely because unaided judgement deviates systematically from normative reasoning under uncertainty [18]. Classical CDSS encoded such principles by design, and it is natural to ask why an automatic reasoner should be expected to rediscover them implicitly, or be fitted to them by training, rather than be equipped with them directly. Second, although clinical reasoning draws on rich, high-dimensional intuition, the search it performs is low-dimensional, discrete, and structured, moving among differentials, tests, and decision steps much as proof search ranges over discrete states [19]; contemporary models, by contrast, expend their effort in a high-dimensional, continuous parameter space, and for all their fluency in generation their search does not traverse the slow, deliberate, discrete structure through which a clinician moves. Third, humans represent and manipulate the world with compositional structures such as decision trees, knowledge graphs, and causal graphs [19], which render reasoning legible and reusable.

These are exactly the properties that classical CDSS provided and that current LLMs lack; the crucial difference is that we no longer wish to hand-craft them. The aim of this thesis is to recover the virtues of the older paradigm, namely explicit, structured, calibrated, and auditable reasoning, without surrendering the generality of learned models, by drawing structure from data and from the knowledge already latent in language models. To this end, the proposal pursues three parallel but connected directions: (i) *separating reasoning from knowledge*, by eliciting the quantities a statistical reasoner needs from a language model, or conversely distilling reasoning structure from the model and populating it with curated data, while inference is carried out in an explicit probabilistic form; (ii) *instilling principled reasoning* into the models themselves, endowing them with decision- and utility-theoretic inference through inference-time procedures and post-training, and grounding this reasoning in learned world models of patient behaviour for robustness and counterfactual analysis; and (iii) *benchmarking* the resulting systems against the reasoning of clinicians and established decision-support tools, using internal clinical decision-tree resources and simulated patients to measure not merely diagnostic accuracy but the alignment, efficiency, and faithfulness of the reasoning itself.

The three works reviewed in the sections that follow trace exactly this trajectory: the reasoning foundations of diagnosis (Section II, Ledley and Lusted [12]), where the standard a trustworthy reasoner should meet is first set; their loss in the move to interactive language models (Section III, MediQ [20]); and a principled but non-clinical attempt to restore them (Section IV, BED-LLM [21]). Section V then develops our research plan, situates our preliminary work, and outlines a timeline for its completion.

II. Reasoning Foundations of Medical Diagnosis

All of this rests on foundations laid, at the dawn of computer-aided medicine, by Ledley and Lusted [12], who set out to analyse how a physician reasons and showed that the process, however intuitive, draws on three mathematical disciplines: symbolic logic, probability, and value theory. The work is the conceptual source of almost everything we ask of a clinical reasoner. Here diagnosis is first written as the application of an explicit body of medical knowledge to a patient's findings, the differential is first given probabilistic meaning, and the choice of treatment is first cast as a decision under uncertainty governed by utilities. It is, in short, where medical reasoning becomes explicitly *reasoning under uncertainty*, the commitment our title inherits, and where the authors fix the standard against which we later judge both classical decision support and the language models that displaced it.

A. The problem

The authors begin from a practical observation: physicians diagnose largely by intuition, forming "a feeling about the case" that is hard to articulate, and the best diagnosticians are simply those who recall and weigh the most possibilities. This is an unreliable foundation, on which errors of omission, never entertaining the right disease, outnumber errors of commission with no systematic guard against them. Their stated motivation is to let computers assist the physician, but they make a prior point that frames the work: before a machine can help, we must understand, in precise terms, *how* the physician reasons. Making the reasoning explicit pays off even without a machine, since it separates routine logical and probabilistic bookkeeping from the genuinely difficult judgements and removes the errors the former breeds. The problem they set is thus to give diagnostic reasoning a mathematical account explicit enough to be analysed, taught, and eventually mechanised.

B. Their solution

The authors decompose diagnosis into three mathematical layers, each supplying what the previous cannot and each separating a body of knowledge from the operation performed on it.

Symbolic logic: Symptoms and diseases are logical attributes and medical knowledge a Boolean function E relating the diseases a patient may have to the findings they may present. A patient presents a symptom complex G; the diagnosis is a function f over diseases; and the whole of diagnosis is the single formula $E \to (G \to f)$, read "if the knowledge E holds, then if the patient presents G, he has diseases f." A "logical basis" enumerating every conceivable disease–symptom combination is pruned by E to a much smaller "reduced basis" of those that can actually occur, which, given a patient's findings, returns the *logically possible* diagnoses, often more than one, and so motivates the next layer.

Probability: Logic narrows the diagnosis to the possible but rarely to one answer, and here the reasoning becomes irreducibly uncertain: a diagnosis, the authors note, "can rarely be made with absolute certainty" and yields only a "most likely" result. Probability manages this residual uncertainty, ranking the possibilities by likelihood. Knowledge enters as conditional probabilities $P(C^k \mid C_i)$, the chance of a symptom complex given a disease complex, with priors $P(C_i)$; Bayes' rule converts these into the posterior $P(C_i \mid C^k)$, the explicit ranked differential. The priors are not universal but local, varying with population, season, and epidemic, and a patient, once diagnosed, joins the statistics on which future diagnoses rest.

Value theory: A diagnosis is not the end; the physician must choose a treatment, a "complicated conflict situation" entangling therapeutic, moral, social, and economic considerations. Drawing on von Neumann's theory of games, the authors score each course of action by its *expected value* and choose the action of highest expected utility, falling back on a minimax mixed strategy when outcome probabilities are unknown. Crucially, this *separates the strategy problem from the decision of values*: it says how to act once the utilities are fixed, leaving the utilities themselves, the moral weights, to human judgement.

C. Results

Being a conceptual paper, it reports no experiments; what it yields is the framework itself, the three pieces of machinery shown to cohere: $E \to (G \to f)$ derives the possible diagnoses, Bayes' rule ranks them by posterior probability, and the expected-utility rule, with a minimax fallback when the probabilities are unknown, chooses the treatment. These are demonstrated only on small, hypothetical cases, not measured on real ones.

Its lasting significance, for us, is that these layers became the blueprint for classical computer-aided diagnosis. The probabilistic pillar in particular was realised at scale in INTERNIST-1 and its probabilistic reformulation QMR-DT [22], [14], which run an inference engine over a hand-curated knowledge base of hundreds of diseases and thousands of findings to compute exactly the explicit posterior differential Ledley and Lusted prescribed. Such systems inherited the foundations' virtues but exposed their price: exact inference over a real knowledge base is intractable, indeed NP-hard, forcing approximation, and the knowledge base must be built and maintained by hand. The foundations are realisable, but classically only over curated knowledge and at steep computational cost.

D. Discussion

This is, to our knowledge, the first precise mathematical account of medical diagnosis, and, decisively for this thesis, the first to separate the *knowledge* a diagnosis draws on from the *inference* performed over it. Both commitments of our title originate here: that diagnosis is *reasoning under uncertainty*, probabilistic because certainty is unattainable, and that knowledge should be held apart from inference.

In the context of our thesis, Ledley and Lusted supply the desiderata by which we judge everything that follows: an explicit probabilistic belief over diagnoses, confidence calibrated to real prevalence, decisions driven by utility rather than likelihood alone, and an inspectable reasoning trace. Classical decision support met them; the turn to language models then bought open-ended fluency at the cost of the explicit belief and the value-driven control, a regression the benchmark we examine next documents [20]. The principled questioning of the work after that [21] restores part of the lost machinery, though outside the clinical, calibrated, utility-aware setting practice demands. Our work returns these foundations to the language-model era.

Read critically, the framework shows its age, and its limits map onto what remains to be done. It is normative and small-scale; it assumes the governing probabilities are known, which the authors concede they "rarely" are; it reduces diseases and findings to crisp Boolean attributes; its logical basis grows combinatorially, the very intractability later systems confront; its hypotheses and findings are fixed in advance, with no way to expand the differential when an unexpected finding appears; its value theory governs only the choice of treatment, not which finding to seek next, leaving diagnosis a one-shot

computation rather than a sequential search for informative evidence; and while it gives a rule for *acting* on utilities, it offers none for calibrating them to real outcomes. Scale, expandable structure, sequential information-gathering, and grounded, calibrated probabilities and utilities are precisely the gaps our work must close; but the standard we are trying to meet was set, with remarkable clarity, here.

III. MediQ: Question-Asking LLMs for Reliable Interactive Clinical Reasoning

We cross now into the language-model era. Classically the foundations could be realised only over hand-built knowledge; language models promise to lift that limit, and the question is whether they keep the reasoning the foundations demand. The first work to put it in clinical terms is MediQ [20]: where the classical account computes a posterior from findings already in hand, a real consultation must first *gather* them. Almost all medical question-answering benchmarks present a case in a single turn, every fact at once, whereas a real consultation begins from incomplete information the clinician must elicit, as earlier automatic-diagnosis dialogue systems also recognised [23]. MediQ recasts diagnosis as this interactive process, in which a competent system must, at each turn, decide whether it knows enough to answer and, if not, ask an informative follow-up; it thereby formalises the information-seeking setting our work targets and, by showing that prompting an LLM alone does not produce this behaviour, locates exactly the gap our explicit belief state and principled questioning rule are meant to fill.

A. The problem

General-purpose LLMs are trained to answer whatever they are asked, returning a plausible response even when the context is incomplete; convenient in low-stakes use, this is dangerous in clinical decision-making, where the model volunteers a confident diagnosis instead of gathering the evidence a clinician would seek. The standard medical-QA paradigm reinforces the habit, presenting a vignette with all findings stated up front and never expecting interaction. Real consultations diverge sharply: patients present partial information and lack the expertise to volunteer every relevant detail, so diagnosis is an investigative process of clarifying questions. A model that scores well on a complete vignette may thus be of little use at the bedside, where the decisive skill is not answering a complete question but recognising an incomplete one. The authors argue for a paradigm shift, from static, single-turn evaluation to an interactive one that begins from incomplete information, and set the challenge of building and measuring systems that know when they have gathered enough to decide, and what to ask when they have not.

B. Their solution

The authors operationalise interactive diagnosis as a dialogue between two language-model systems, with the apparatus to evaluate each.

The interactive task: The full record contains a subset of facts sufficient for the diagnosis, but the system is given only an opening presentation (age, sex, chief complaint) and must expand its knowledge turn by turn until that subset is covered, then answer a multiple-choice question. Performance is scored on the accuracy of the final decision and the efficiency (number of follow-ups) of the exchange.

The Patient system: One system holds the record and answers, judged on factuality (faithful to the record) and relevance (addressing the question). Of three variants (Direct, Instruct, and Fact-Select), the last, which decomposes the record into atomic facts and answers by selecting among them rather than generating free text, is markedly the most reliable, an early sign (which our own preliminary system exploits) that reasoning over discrete, atomic evidence reduces hallucination relative to raw text.

The Expert system: The other, which the authors call MediQ-Expert, plays the clinician through five modular steps: an initial assessment naming likely knowledge gaps; an abstention module deciding whether to answer or ask; question generation; information integration appending each answer to the history; and a final decision.

Eliciting the abstention decision: The crux is abstention, deciding when to ask rather than answer, which the authors probe with graded confidence-elicitation strategies: a basic ask-or-answer prompt; a numerical score thresholded at a cut-off; a binary sufficiency judgement; a five-point scale; and, layered on these, rationale generation and self-consistency over repeated samples, with the action thresholds chosen by grid search.

Building the testbed: A reusable recipe converts any static medical benchmark into a multi-turn one (reduce the case to its initial presentation and re-serve the rest only on request), yielding iMedQA and iCraft-MD from MedQA and the dermatology set Craft-MD, on which they evaluate Llama-3-Instruct (8B and 70B), GPT-3.5, and GPT-4.

C. Results

The results are pointed. Accuracy rises monotonically with information available (None to Initial to Full), confirming that information seeking, not raw reasoning, is the binding constraint; yet simply instructing a model to ask questions *lowers* accuracy by 11.3% relative to answering from the same limited information, and in practice the models barely ask, GPT-3.5 posing on average 0.47 questions per case. The Fact-Select Patient, by contrast, anchors the benchmark at 89.1% factuality and 79.9% relevance. Only the richest abstention strategy, a five-point scale with rationale generation and self-consistency, reverses the failure to ask, improving on the question-asking baseline by 22.3% and on the non-interactive baseline by 12.1%; even so, a 10.3% gap to the full-information upper bound remains, and the best configuration closes only 51.2% of the partial-to-full distance. The benefit is confined to the largest models: only those above seventy billion parameters surpass the non-interactive baseline, while smaller ones are made *worse* by asking. A finer analysis shows rationale generation helps mainly by improving *which* question is asked rather than *when*, sharpening calibration from an expected error of 0.286 to 0.211, with self-consistency helping

only alongside it. Even the strongest current models, then, cannot yet reliably judge when they know enough or what to ask next.

D. Discussion

MediQ matters to us in three ways and is bounded in a fourth. It formalises the interactive clinical setting our work targets and leaves a reusable benchmark that directly informs the reasoning-aware evaluation of our third direction; its abstention module is an operational form of the know-when-to-defer requirement trustworthy support demands; and its central negative result is, for us, motivating rather than discouraging, locating exactly the gap an explicit belief state and a principled questioning rule must fill. It is, moreover, an early and specifically medical instance of a failure now broadly documented, in which otherwise capable models lose the thread of multi-turn exchanges [10]. Its limits map onto our directions: confidence and abstention are elicited by prompting rather than read from a calibrated posterior, so the quantity that should govern the decision to ask is never represented explicitly; the thresholds are grid-searched and reported at their best, which flatters the method; no value-of-information criterion governs *which* question to ask, a rule supplied by the work we examine next [21]; the inference stays inside the model and is not auditable; the task is multiple-choice rather than open-ended diagnosis with real prevalence, utilities, and calibrated deferral; and the simulated patient, itself a language model only about ninety per cent factual, leaks its errors into the benchmark; and it scores only the final answer and the number of questions, never the soundness of the reasoning that produced them, so a right answer reached for the wrong reasons still counts as success, the process-level gap our third direction sets out to measure. MediQ thus establishes the problem and proves that prompting alone will not solve it; the machinery that would, an explicit calibrated belief, a principled rule for what to ask, and auditable inference, is what the thesis sets out to supply.

IV. BED-LLM: Intelligent Information Gathering with LLMs and Bayesian Experimental Design

MediQ marks the gap; the natural question is how to close it, and our last reviewed work offers one concrete example. BED-LLM [21] shows how the principled, value-driven rule MediQ lacks can be built directly on a language model. It treats adaptive, multi-turn information gathering with an LLM as sequential Bayesian experimental design (BED): at each turn it chooses the question of greatest expected information gain (EIG) about an unknown of interest, then updates its beliefs on the answer. This is the value-of-information criterion our work requires, computed explicitly rather than left to the model's discretion, and over a small, discrete space of candidate questions rather than the high-dimensional continuum a model ordinarily searches, which makes BED-LLM the closest existing counterpart to the questioning engine at the centre of the reasoner we propose: it draws its knowledge (the candidate hypotheses and the likelihood of answers) from the LLM while keeping the inference that selects the next question explicit and auditable.

A. The problem

Despite their successes, LLMs are poor at proactively seeking information. Asked to clarify a request, play a guessing game, or elicit preferences, a modern LLM can produce a coherent single question but struggles to *tailor* its questions to answers already gathered: it repeats itself, asks what the dialogue has settled, or commits prematurely to one guess. This matters wherever a model must gather evidence before acting (clarifying intent, running a survey, or taking a clinical history), because it is not enough to generate good questions up front; the next question must depend on the answers so far. Of the two ways to improve this, fine-tuning the model or changing how it is used at deployment, the authors take the latter: deployment-time methods need no task-specific data up front (a user's preferences cannot be collected in advance) and apply to any existing LLM. The challenge is thus to turn a fixed, pre-trained LLM into a principled, adaptive questioner.

B. Their solution

The authors cast the task as sequential BED and instantiate every piece of the framework from the LLM. Their method iterates a five-step loop each turn: extract a belief over the unknown; generate diverse candidate questions; estimate each one's EIG; ask the highest-EIG question; and update the history with the answer. The design decisions that make this loop work are as follows.

The information-gain objective: EIG is the mutual information between the unknown θ and the answer y a question x would elicit, equivalently the expected reduction in the entropy of the belief over θ once the answer is known,

$$\mathrm{EIG}(x) = I(\theta; y \mid x) = H[p(\theta)] - \mathbb{E}_y[H[p(\theta \mid y, x)]]\,; \quad (1)$$

maximising it selects the most informative question. Because the objective depends only on the current model, it updates automatically as the belief is refined, which suits a turn-by-turn design.

Constructing the model: The only model at hand is the LLM, so the authors pair a prior over hypotheses $p(\theta)$ with an LLM-derived likelihood $p_{\mathrm{LLM}}(y \mid \theta, x)$. They contrast this with a data–estimation pairing that simulates answers and infers θ, and argue, analytically and empirically, for the prior–likelihood form when the answer space is simpler than the hypothesis space, as here; it also lets the belief over θ be read off directly, as valid sequential BED requires.

Belief construction and filtering: A central finding is that updating beliefs by in-context prompting fails: even strong LLMs sample hypotheses incompatible with past answers and grow overconfident as the dialogue lengthens. The authors instead sample hypotheses at raised temperature for diversity, then *filter* them: each is scored against every prior question–answer pair and rejected if its likelihood of the observed answers falls below a fixed threshold, with survivors retained across turns so the belief is refined rather than rebuilt. The filter cheaply reimposes the discipline exact Bayesian conditioning would enforce but that would be prohibitive to run through the LLM each turn.

Estimating the EIG: Because the answer space is far smaller than the hypothesis space, the same quantity is rewritten, by the symmetry of mutual information, over the *answer*,

$$\mathrm{EIG}(x) = H[p(y \mid x)] - \mathbb{E}_\theta[H[p(y \mid \theta, x)]]\,, \qquad (2)$$

the entropy of the predicted answer marginalised over the surviving hypotheses minus its average entropy conditioned on each. The first term rewards questions whose answer is uncertain overall; the second penalises those whose answer stays uncertain even once the hypothesis is fixed and so cannot discriminate. Because the question space cannot be searched directly, the candidates are themselves proposed by the LLM, sampled freely from the history or conditioned on the surviving hypotheses to split them, and confined to a small multiple-choice answer set so the entropies are meaningful.

C. Results

In the 20 Questions game, where the system identifies a hidden entity through up to twenty yes/no questions across three sets (Animals, Celebrities, Things) of 100 entities and six LLMs, BED-LLM raises the final success rate by 37.4 percentage points over direct prompting; the gain is positive in every model–category combination, never decreases over the game, and more than doubles the baseline in over half the setups (e.g. GPT-4o-mini on Animals from 44% to 88%, Mistral-Large from 33% to 95%). It beats the previous state of the art, and the authors' ablations confirm that the gain comes from the ingredients we care about: the full information criterion rather than a partial one and, decisively, the filtered belief, since forming beliefs by in-context prompting alone collapses performance. The advantage persists under a questioner–answerer mismatch, where simpler baselines turn brittle, and a second task, inferring a user's film preferences over a few turns on 200 simulated profiles, repeats the pattern. The lesson: a principled criterion over a disciplined belief extracts far more from each turn than instructing a model to "ask a good question."

D. Discussion

This is, to our knowledge, the first application of sequential BED to interactive information gathering with LLMs without a deterministic-likelihood assumption, and its careful treatment of the joint-model factorisation, the belief update, and the EIG estimator is its contribution. For us it is valuable because it instantiates, over open-ended language, the value-driven questioning classical decision support could only express over a hand-built knowledge base: the hypotheses and their likelihoods come from the LLM while the rule for choosing the next question stays explicit and auditable. Its experimental-design heritage also points at our second direction, since the same EIG objective can be amortised into a learned questioning policy.

Our main critique is the distance from here to clinical decision support. The demonstrations are non-clinical (a parlour game and a recommendation task), and the evaluation runs entirely between language models, one answering another and, in the preference task, a model judging the result, so the gains are never checked against a human or a real outcome. The belief is uniform over whatever hypotheses survive filtering, so its calibration is only as trustworthy as the model that proposes them and is never anchored in real epidemiology, while any hypothesis the sampler fails to generate is unrecoverable, quietly bounding the claim to an unrestricted hypothesis space. The objective is myopic, choosing the single most informative next question with no lookahead, and it selects by information alone, with no notion of cost or risk, so it cannot weigh an informative but dangerous test against a safer one, or decide when to stop and defer, the expected-utility reasoning the founding analysis [12] demanded and our reasoner must restore. It is, moreover, wholly unamortised and so costly at deployment: each turn re-samples and re-filters the hypotheses and scores every candidate question through many forward passes of the model, a price that recurs at every step with nothing learned across turns or cases. These are exactly the gaps our work closes: we ground the questioning engine in clinical structure, anchor its beliefs in real data, and carry its questions through to a calibrated, utility-aware decision.

V. Research Proposal

Expert clinicians reason under uncertainty in a characteristic way: they hold several diagnostic hypotheses at once, seek the evidence that most reduces their uncertainty, act only when the balance of risks justifies it, defer when it does not, and can account for each step afterwards. The central question of this thesis is how to bring this mode of reasoning to language-model-based systems for high-stakes clinical decision-making, whether by using language models as components within principled reasoning systems, or by instilling principled reasoning into the models themselves. We pursue it through three parallel but interconnected directions: separating reasoning from knowledge, instilling principled reasoning into the models, and measuring how faithfully the resulting systems reason as clinicians do. Together they aim at a concrete contribution that consolidates what the programme learns: *Meditron Reasoning*, an extension of our lab's open Meditron models [24] into a clinical reasoner that holds a calibrated belief over diagnoses, gathers evidence by its value, and abstains or defers when it does not know enough to act safely. Figure 2 lays out the four-year plan across these directions.

Our preliminary work establishes the premise on which the programme rests. MoBayes [25] is a modular system in which the language model serves only as a linguistic interface, translating free-form patient dialogue into structured clinical observations, while a separate probabilistic module maintains an explicit posterior over diagnoses, chooses follow-up questions by their expected information gain, and decides when to commit or defer under calibrated thresholds. This separation outperforms standalone frontier diagnostic systems [2]; strikingly, even a knowledge base elicited from the language model itself, once placed inside the probabilistic engine, is used more effectively than by the same model reasoning end-to-end. The same separation lets an inexpensive model paired with the engine outmatch far larger autonomous systems at a fraction of the cost, holds up under evasive, adversarial

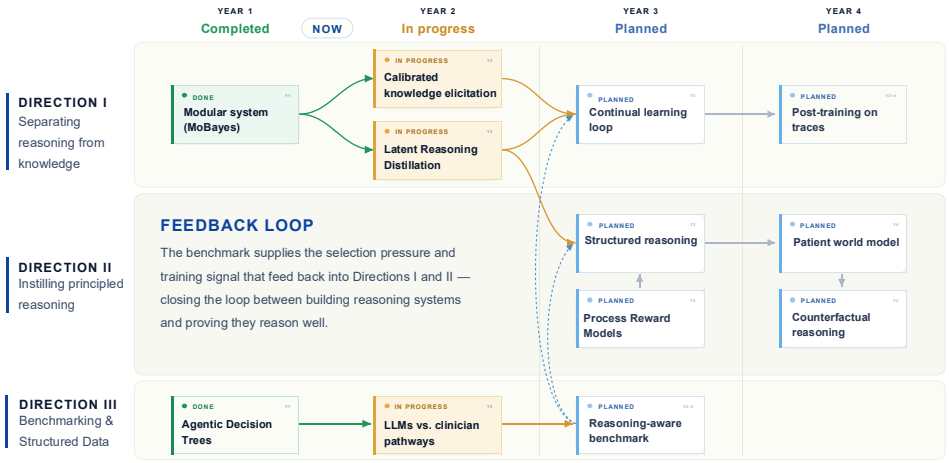

Fig. 2. Doctoral research roadmap. The three directions run in parallel across four years; horizontal position encodes time, arrows show dependencies, and the reasoning-aware benchmark of Direction III feeds back into Directions I and II.

patient communication, and adapts to a new population by swapping its statistical backend rather than retraining (Fig. 3). Two properties of this design matter for what follows: because the language model never asserts a probability of its own, calibration is enforced by the engine rather than hoped for from the model; and because each step runs through an explicit module, the whole trace, from parsed observation through chosen question to final decision, can be inspected and audited rather than reconstructed after the fact. The result is encouraging but provisional, since it rests on assumptions that the three directions are designed to relax: that knowledge can be extracted once and trusted, that the reasoning structure is fixed and external, and that quality is captured by final accuracy.

A. *Direction I: Separating reasoning from knowledge*

The first direction asks whether the two ingredients of clinical reasoning, what one knows and how one reasons with it, can be cleanly separated and recombined, and it admits two complementary readings. In the first, the language model is treated as a *repository of knowledge*: we seek principled and reliable ways to elicit the quantities a statistical reasoner requires, namely the candidate conditions, their prevalence, and the likelihoods relating findings to conditions, and to perform inference over them in an explicit probabilistic model, as a recent line of work does in a single step [8] and as classical diagnostic networks did over hand-specified knowledge [22]. The open problems are the faithfulness and calibration of what is extracted, and its scope: rather than harvesting an entire knowledge base in advance, the system can use its current beliefs to elicit only the quantities worth acquiring next, turning the informal aim of extracting the right knowledge into a precise, value-driven query. In the second reading, the language model is treated instead as a source of *reasoning structure*: we ask whether the compositional scaffolds over which diagnosis proceeds, the differential hierarchies, finding-condition graphs, and decision procedures, can be generated, or the model's latent logic distilled, and then populated with curated real-world statistics in place of the model's own estimates. Such scaffolds need not be conjured from nothing: the corpus of clinical decision algorithms described in Direction III is itself an explicit, expert-authored instance of this structure, which a language model may extend rather than reinvent.

These readings converge in a more ambitious formulation in which the language model and the statistical modules operate together in a continual loop. The model consults the modules to reason and decide, but also recognises where the current structure is inadequate, whether an unanticipated finding, a missing dependency, or a hypothesis outside the differential, and expands the knowledge and compositional structures accordingly, so that representation grows through use rather than being frozen at design time. The structures and reasoning traces this loop produces are, in turn, a natural training signal: a further extension uses them to post-train the model so that it comes to perform such structured, calibrated reasoning natively, rather than depending on an external engine at every step.

B. *Direction II: Instilling principled reasoning*

Where the first direction composes language models with external reasoners, the second, more exploratory one asks whether the principles themselves can be carried inside the model. Drawing on statistical decision theory, utility theory, and probability, we will study inference-time scaling strategies that lead a model to maintain explicit beliefs, to value candidate questions and tests by their expected information, and to choose actions by expected utility, with the reasoning rendered as an inspectable trace rather than concealed in activations; and we will ask whether this behaviour can be made permanent through post-training, for which our decision-tree corpus supplies a natural process reward model, scoring reasoning steps for their consistency with the guideline pathway. Test-time and agentic paradigms offer a complementary route for instilling particular reasoning modes, among them abductive hypothesis formation, within the model's own deliberation.

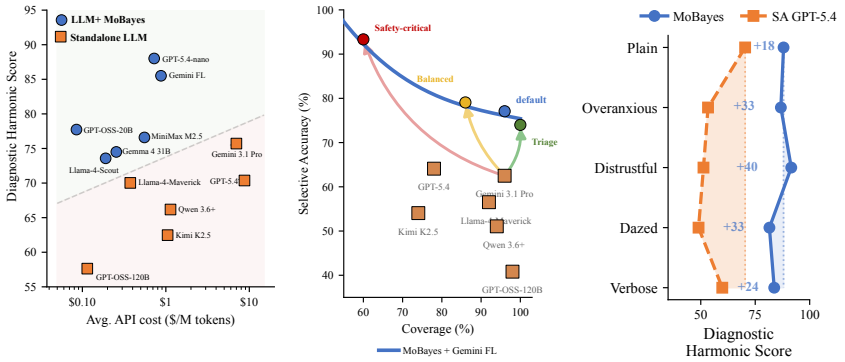

Fig. 3. Preliminary results from our modular system, MoBayes [25]: *(a)* on the cost–accuracy plane, cheap models inside the engine (circles) dominate standalone frontier doctors (squares) at far lower cost; *(b)* a controllable accuracy–coverage frontier (answer vs. defer); *(c)* robustness across adversarial patient personas.

A natural methodological tool for the learned half of this agenda is amortised inference over structured objects: learning to *sample* compositional structures in proportion to their value rather than collapsing onto a single answer [26], [27]. This property is well matched to diagnosis, where several differentials must be kept alive in proportion to their probability, and it has already been used to fine-tune language models to sample latent reasoning chains from a posterior [28]. We adopt such tools as means to an end, namely calibrated, diverse, structured reasoning, rather than as objects of study in themselves. Robust reasoning must finally be grounded in how patients actually behave, not in idealised vignettes, and this grounding is a central aim of the direction. We will learn *world models* of patients from electronic health records [29], generative models of patient trajectories that let the reasoner rehearse a case against realistic dynamics and ask *counterfactual* questions, how a patient would have evolved under a different action, the comparison a clinician weighs when choosing between options. By recovering the *causal* structure underlying these trajectories, the system comes to reason about interventions rather than mere associations and stays robust under the distribution shifts of real practice, turning the decision-theoretic machinery above into beliefs, information values, and utilities that can be tested against how the patient and the disease actually respond.

C. Direction III: Benchmarking clinical reasoning

The preceding directions make a claim about *process*, that these systems reason as clinicians do, yet the field evaluates almost exclusively on final diagnostic accuracy. The third direction develops the means to measure reasoning itself. We have access, within our lab, to EPOCT+, a curated corpus of structured clinical decision algorithms on the order of 10^5 decision trees that encode the pathways clinicians and established decision-support systems follow, against which a system's reasoning can be compared. Using simulated patients derived from these algorithms and from learned trajectory models, we will define metrics that quantify how closely a system's reasoning aligns with the clinician pathway and with classical decision-support output: the divergence between its question-selection policy and the guideline, the efficiency with which it reaches calibrated confidence, and the faithfulness of its stated reasoning to its actual computation. The outcome is a reasoning- and calibration-aware benchmark grounded in real clinical practice, which not only evaluates the systems of Directions I and II but supplies the selection pressure and training signal that feed back into them, closing the loop between building reasoning systems and establishing that they reason well.

These resources extend beyond the decision trees. MOOVE, a second initiative the lab maintains, records expert preferences among management options across clinical scenarios, giving direct supervision for the utilities and value-of-information judgements that Directions I and II rely on. Looking ahead, four prospective clinical trials over the next three years, in Switzerland, the United States, and Tanzania, among other African countries, will collect the real-world data with which to train and stress-test the patient world models of Direction II and to benchmark the resulting reasoners against practice.

D. Timeline

The three directions run in parallel, with dependencies that stagger their emphasis across the doctorate, as Figure 2 sets out. The work to date has delivered the modular system above and assembled the structured decision-tree resources on which the benchmark rests. The current year develops Direction I's calibrated, demand-driven knowledge elicitation and its distillation of reasoning structure, alongside the first benchmarking of language models against clinician pathways in Direction III. The third year turns to the continual, structure-expanding loop of Direction I and to Direction II, where the decision-tree corpus drives a process reward model and the information objective is amortised into a learned policy, all measured by the maturing reasoning-aware benchmark. The fourth year post-trains the model on the traces this produces and grounds it in a record-derived patient world model that supports counterfactual reasoning, before consolidating the three threads into a single account of trustworthy clinical reasoning under uncertainty. Foundations from decision and utility theory, value of information, and calibrated probability estimation [6], [7] run throughout.

REFERENCES

[1] D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, and P. Szolovits, "What disease does this patient have? a large-scale open domain question answering dataset from medical exams," *Applied Sciences*, vol. 11, no. 14, p. 6421, 2021.

[2] T. Tu, A. Palepu, M. Schaekermann, K. Saab, J. Freyberg, R. Tanno, A. Wang, B. Li, M. Amin *et al.*, "Towards conversational diagnostic AI," *Nature*, vol. 642, no. 8067, pp. 442–450, 2025.

[3] D. McDuff, M. Schaekermann, T. Tu, A. Palepu, A. Wang, J. Garrison, K. Singhal, Y. Sharma, S. Azizi *et al.*, "Towards accurate differential diagnosis with large language models," *Nature*, vol. 642, no. 8067, pp. 451–457, 2025.

[4] L. Yang, S. Xu, A. Sellergren, T. Kohlberger *et al.*, "Advancing multimodal medical capabilities of Gemini," *arXiv preprint arXiv:2405.03162*, 2024.

[5] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, "Large language models in medicine," *Nature Medicine*, vol. 29, no. 8, pp. 1930–1940, 2023.

[6] M. Xiong, Z. Hu, X. Lu, Y. Li, J. Fu, J. He, and B. Hooi, "Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs," in *International Conference on Learning Representations (ICLR)*, 2024.

[7] B. Li, B. Zhou, F. Wang, X. Fu, D. Roth, and M. Chen, "Deceptive semantic shortcuts on reasoning chains: How far can models go without hallucination?" in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Mexico City, Mexico: Association for Computational Linguistics, 2024, pp. 7675–7688. [Online]. Available: https://aclanthology.org/2024.naacl-long.424/

[8] Y. Feng, B. Zhou, W. Lin, and D. Roth, "BIRD: A trustworthy bayesian inference framework for large language models," in *International Conference on Learning Representations (ICLR)*, 2025.

[9] O. Liu, D. Fu, D. Yogatama, and W. Neiswanger, "DeLLMa: Decision making under uncertainty with large language models," *arXiv preprint arXiv:2402.02392*, 2024.

[10] P. Laban, H. Hayashi, Y. Zhou, and J. Neville, "LLMs get lost in multi-turn conversation," in *International Conference on Learning Representations (ICLR)*, 2026.

[11] D. Fan, S. Delsad, N. Flammarion, and M. Andriushchenko, "HalluHard: A hard multi-turn hallucination benchmark," *arXiv preprint arXiv:2602.01031*, 2026.

[12] R. S. Ledley and L. B. Lusted, "Reasoning foundations of medical diagnosis," *Science*, vol. 130, no. 3366, pp. 9–21, 1959.

[13] F. T. de Dombal, D. J. Leaper, J. R. Staniland, A. P. McCann, and J. C. Horrocks, "Computer-aided diagnosis of acute abdominal pain," *British Medical Journal*, vol. 2, no. 5804, pp. 9–13, 1972.

[14] R. A. Miller, H. E. Pople, and J. D. Myers, "INTERNIST-1, an experimental computer-based diagnostic consultant for general internal medicine," *New England Journal of Medicine*, vol. 307, no. 8, pp. 468–476, 1982.

[15] J. von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior*. Princeton University Press, 1944.

[16] R. D. Luce and H. Raiffa, *Games and Decisions: Introduction and Critical Survey*. Wiley, 1957.

[17] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. Springer, 1985.

[18] M. H. Bazerman and D. A. Moore, *Judgment in Managerial Decision Making*, 8th ed. Wiley, 2012.

[19] E. J. Hu, "Building a reasoning machine," Ph.D. dissertation, Université de Montréal, 2026.

[20] S. S. Li, V. Balachandran, S. Feng, J. S. Ilgen, E. Pierson, P. W. Koh, and Y. Tsvetkov, "MediQ: Question-asking LLMs and a benchmark for reliable interactive clinical reasoning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[21] D. Choudhury, S. Williamson, A. Goliński, N. Miao, F. Bickford Smith, M. Kirchhof, Y. Zhang, and T. Rainforth, "BED-LLM: Intelligent information gathering with LLMs and bayesian experimental design," in *International Conference on Learning Representations (ICLR)*, 2026.

[22] M. A. Shwe, B. Middleton, D. E. Heckerman, M. Henrion, E. J. Horvitz, H. P. Lehmann, and G. F. Cooper, "Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base. I. the probabilistic model and inference algorithms," *Methods of Information in Medicine*, vol. 30, no. 4, pp. 241–255, 1991.

[23] L. Xu, Q. Zhou, K. Gong, X. Liang, J. Tang, and L. Lin, "End-to-end knowledge-routed relational dialogue system for automatic diagnosis," in *AAAI Conference on Artificial Intelligence*, vol. 33, no. 1, 2019, pp. 7346–7353.

[24] Z. Chen, A. Hernández Cano, A. Romanou *et al.*, "MEDITRON-70B: Scaling medical pretraining for large language models," *arXiv preprint arXiv:2311.16079*, 2023.

[25] Y. Kesmen, F. Elhassan, J. Ma, J. Stalhandske, Y. Chang, D. Sasu, A. Kulinkina, A. Arora, L. Klein, and M.-A. Hartley, "MoBayes: A modular bayesian framework for separating reasoning from language in conversational clinical decision support," 2026.

[26] E. Bengio, M. Jain, M. Korablyov, D. Precup, and Y. Bengio, "Flow network based generative models for non-iterative diverse candidate generation," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, 2021, pp. 27 381–27 394.

[27] N. Malkin, S. Lahlou, T. Deleu, X. Ji, E. J. Hu, K. Everett, D. Zhang, and Y. Bengio, "GFlowNets and variational inference," in *International Conference on Learning Representations (ICLR)*, 2023.

[28] E. J. Hu, M. Jain, E. Elmoznino, Y. Kaddar, G. Lajoie, Y. Bengio, and N. Malkin, "Amortizing intractable inference in large language models," in *International Conference on Learning Representations (ICLR)*, 2024.

[29] M. A. Qazi, M. Nadeem, and M. Yaqub, "Beyond generative AI: World models for clinical prediction, counterfactuals, and planning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.