COTCAGENT: PREVENTIVE CONSULTATION VIA PROBABILISTIC CHAIN-OF-THOUGHT COMPLETION

Anonymous authors
Paper under double-blind review

ABSTRACT

Sequential diagnosis requires ranking diseases from a sparse intake under a public budget of follow-up questions, which is diagnostic triage with missing findings rather than screening of people who have no symptoms. Because laboratory series arrive irregularly and histories remain incomplete, the ranking must be updated as new facts arrive, yet most pipelines are limited by unnamed numeric trends, unverifiable free-form chain of thought, and a negative bias from treating unasked findings as absent. To solve these, we propose COTCAgent, a sparse-intake consultation agent built on Chain-of-Thought Completion (COTC), which couples three modules on one shared ternary evidence log. The Temporal-Statistics Adapter (TSA) turns time series into typed predicates with named statistics, while the COTC module scores diseases with a learnable knowledge-base energy in which unknown findings add zero energy, after which bounded probabilistic completion selects unknown findings under the question budget and writes each answer back as present, absent, or unknown. Edge strengths and an absence scale are trained with listwise ranking under a sparse curriculum, so that the same energy chooses the next finding by a surrogate entropy reduction. Experiments on DDXPlus and MIMIC-IV show that these modules improve ranking under a matched question budget, with information-gain completion above asking nothing or at random and above protocol-matched askers and a local LLM, thereby providing a foundation for future calibrated scoring and prospective preventive consultation.

1 INTRODUCTION

Sequential diagnosis is hard because labs arrive at irregular times, histories are incomplete, and the ranked list of diseases must change as new facts arrive Zhou et al. (2025); Wornow et al. (2024); Nori et al. (2025). In sequential diagnosis, preventive consultation ranks diseases from a sparse intake: the agent may ask a public number of follow-up questions and then returns a ranked list. The setting is triage with missing findings, not screening of people with no symptoms. DDXPlus cases and MIMIC-IV admissions are already symptomatic or already in hospital. The practical questions are which unknown finding to ask next so that the ranked list improves, and how each update can be checked against a named computation rather than against free-form text.

Language models are widely used for medical dialogue, monitoring, and tool-using workflows Singhal et al. (2025); Liu et al. (2024); Qin et al. (2024). Three limitations recur. First, numeric trends are often written in words with no named estimator, so slopes and change points are hard to check. Second, free-form chain of thought can draft questions, but the resulting trace is hard to check against a structured knowledge base. Third, many interactive pipelines turn incomplete charts into full symptom vectors and mark every unasked finding as absent. That pattern appears in older tabular classifiers and RL symptom checkers, and also in LLM stacks that fill unmentioned findings as negative. The result is a strong negative bias, a penalty on incomplete intakes, and little value to asking.

Medical LLMs were often tested on templated intakes or multiple-choice items that do not look like chart review Singhal et al. (2025); Saab et al. (2024). Longer context windows change how much of a record fits in a prompt Liu et al. (2025); Wornow et al. (2024). They do not by themselves give faithful numeric reasoning or questioning under a budget. Tool-using and multi-agent stacks move arithmetic and retrieval out of the decoder Liu et al. (2024); Luo et al. (2025); Qin et al. (2024).

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

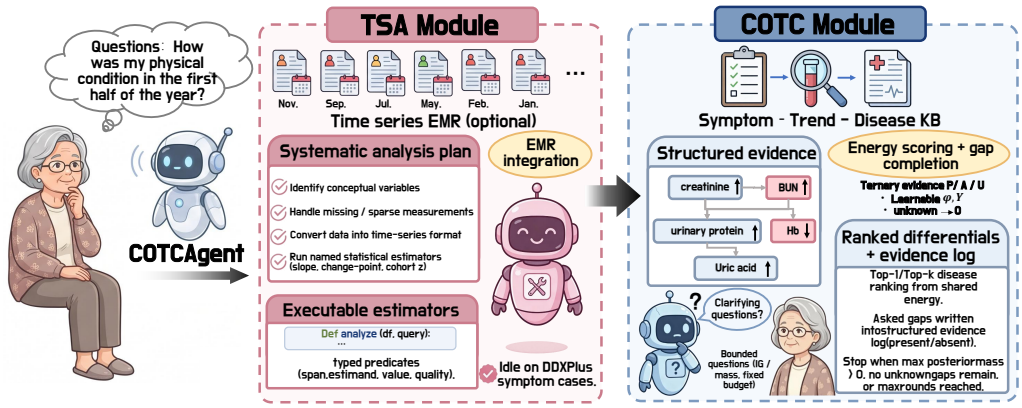


Figure 1: COTCAgent overview. TSA emits typed trend predicates from time series when those series exist. The COTC module scores diseases and drives bounded completion over ternary evidence.

Agents that read long records can still invent trend claims and rationales that are hard to map back to knowledge-base rows Feli et al. (2025). We therefore focus on preventive consultation under a matched question budget: complete a structured evidence log, re-rank diseases with a score that can be replayed, and report both Hit@1 and question cost so that policies can be compared.

We do not optimize open-domain dialogue or free-form clinical notes. We ask whether a small knowledge-grounded scorer can turn sparse intakes into stronger ranked lists when every follow-up question is counted and every answer maps into a finite evidence vocabulary.

We propose COTCAgent in Section 3 and Algorithm 1 as three layers, shown in Figure 1. The Temporal-Statistics Adapter turns time series into typed trend predicates with executable estimators when such series exist. The Chain-of-Thought Completion module, abbreviated COTC, scores diseases with inverse-disease-frequency weights and a learnable edge matrix ϕ over a symptom, trend, and disease knowledge base. This module is the only part trained from labeled differentials. Bounded probabilistic CoT completion then asks only about unknown findings, parses short answers into ternary evidence, and stops when a mass threshold is met, when no unknown finding remains, or when the round limit is reached. Chain of thought here is the growing evidence log and the energy trail, not token-level reasoning inside a language model. The loop is iterative completion over a finite vocabulary. We restrict the agent to templated questions so that each step stays tied to a knowledge-base row and a dialog turn. Our contributions can be summarized as follows:

- We propose COTCAgent, a sparse-intake consultation agent that couples executable TSA predicates, a learnable COTC scoring module, and bounded probabilistic CoT completion on one shared ternary evidence log. Unknown findings add zero energy, and present or absent findings update a knowledge-base energy that can be replayed.
- We introduce listwise training of edge strengths ϕ and an absence scale γ under a sparse curriculum. The same trained energy is reused for ranking and for follow-up selection, so each update stays tied to a knowledge-base row rather than to free-form text.
- We evaluate matched-budget consultation on DDXPlus and MIMIC-IV. Information-gain completion reaches 88.1% Hit@1 with 3.61 questions on DDXPlus and lifts phenotype Hit@1 from 72.7% to 79.4% on MIMIC-IV. Protocol-matched copies of recent acquisition methods and a local LLM that both asks and ranks stay below learned COTC under a shared answer oracle.

2 Related Work

Medical LLMs, Tool Agents, and Sequential Diagnosis. Medical LLMs have moved from static QA and capability tests Singhal et al. (2025); Saab et al. (2024); Esteva et al. (2019) toward long-context EHR modeling Wornow et al. (2024); Liu et al. (2025). Agent benchmarks move planning, retrieval, and code outside the decoder Liu et al. (2024); Qin et al. (2024); Luo et al. (2025). Multi-agent

stacks widen collaborative diagnosis Qian et al. (2024). Sequential protocols stress coordinated questioning Nori et al. (2025); Zhou et al. (2025). Many systems still mix drafting, evidence updates, and ranking in free-form generation. Cost and checkability are then hard to see under sparse intakes. COTCAgent keeps language at template questions and short-answer parsing, so the COTC module stays an explicit energy with a public question cost. We compare against a local LLM that both asks and ranks in Table 3.

Knowledge-Grounded and Neuro-Symbolic Medical Reasoning. A second line grounds predictors in graphs or rules Xu et al. (2025); Wang et al. (2024); DeLong et al. (2023); García-Barragán et al. (2025). Symptom checkers often use fixed likelihood tables. Those tables are easy to inspect but hard to adapt when edges should be learned from labels. On DDXPlus, BASD, AARLC, and Diaformer are standard interactive baselines Tchango et al. (2022); Yuan & Yu (2024); Xu et al. (2019); Chen et al. (2022). COTCAgent learns continuous edge strengths on a discrete incidence matrix, keeps unknown slots explicit, and does not invent knowledge-base relations when no unknown finding remains.

Probabilistic Scoring, Missingness, and Active Acquisition. Graphical models give a clear meaning when priors exist Koller & Friedman (2009). Many ranked lists still treat unseen findings as absent. Ternary present, absent, and unknown logs are already standard in interactive symptom checkers Tchango et al. (2022); Yuan & Yu (2024). We keep that encoding so unknown findings add zero energy under listwise training. Expected information gain is classical experimental design Guan & Baral (2021). Because our scorer is discriminative rather than generative, we use a one-step information-gain surrogate on the trained ranking scores $\tilde{P}$, not calibrated mutual information from a symptom likelihood and not token-level LLM uncertainty. Each update writes a ternary token onto a knowledge-base incidence row and adds to a public question cost. We compare information-gain completion to mass, frequency, fixed-KB BED, and learned neural askers under the same answer oracle.

3 METHOD

3.1 OVERVIEW OF COTCAGENT

COTCAgent ranks diseases with optional follow-up questions, as shown in Figure 1. TSA produces predicates from labs and vitals when time series exist. The COTC module matches symptoms and predicates against a knowledge base and returns a ranked list with explicit unknown findings, as in Figure 2. Bounded completion fills those unknown findings until a stopping rule fires. Natural language is used only to render templated questions and to parse short answers into discrete evidence.

We keep three surfaces separate so that failures stay easy to see. Executable statistics consume raw series under named likelihoods, so slope and change-point claims are regressions rather than free-text guesses. The COTC module consumes only predicates and evidence tokens that belong to a finite schema. Natural language only paraphrases those pre-specified unknown findings and parses short answers back into the same tokens. When TSA fails, the COTC module receives an explicit uncertainty flag instead of an invented numeric claim. When no unknown finding remains or the mass threshold is met, Algorithm 1 returns the current ranking and does not invent knowledge-base edges.

3.2 TEMPORAL-STATISTICS ADAPTER

TSA maps a clinical analysis request Q to typed predicates through a deterministic adapter

$$f_{\mathrm{TSA}} : Q \to \mathcal{I} \to \Lambda \to C, \tag{1}$$

where $\mathcal{I}$ is a rule-parsed intent (which series and which target statistic), Λ is a named estimator family, and C is the executable routine that implements that family. The request Q is not free-form user chat. It is a structured tuple of series name, timestamps, values, and a target key such as slope or change-point. The planner inside TSA is a lookup from that key to a routine. It is not a language-model planner, and it does not generate code at evaluation time. Trend queries fit an ordinary least-squares slope on the current patient's series,

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + u_i + \epsilon_{ij}, \qquad u_i \sim \mathcal{N}(0, \sigma_u^2),\ \epsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon^2), \tag{2}$$

where i indexes the patient, j the observation, and t_{ij} time. Equation 2 records the cohort-level random-intercept extension for settings in which repeated patients are modeled jointly; the online single-patient TSA uses the OLS special case $u_i = 0$ and does not claim to estimate σ_u^2 from one trajectory. When a training cohort is available, as on PhysioNet, TSA also attaches a cohort z-score that uses the training-split mean and variance of the last observation. Abrupt-change queries scan a split index τ by reduction in residual sum of squares,

$$\text{score}(\tau) = \text{SSE}(\mathbf{y}) - \big(\text{SSE}(\mathbf{y}_{1:\tau}) + \text{SSE}(\mathbf{y}_{\tau+1:T})\big), \qquad \tau \in \{2, \dots, T-2\}, \tag{3}$$

which matches a Gaussian two-piece mean-shift model with a uniform prior on τ. Failed fits fall back to simpler diagnostics and attach SPARSE or UNSTABLE flags. Instability is residual RMSE relative to the series mean, so the COTC module can down-weight the predicate. Each TSA artifact is a typed tuple of span, target statistic, value, and quality drawn from a finite schema that covers slope, change-point score, and cohort z. Residual magnitude is a quality flag, not a separate evidence token. Numeric claims stay tied to named routines. The planner only selects which target statistic to run, and the returned object is always the output of that documented procedure.

On DDXPlus, the temporal estimators are unnecessary because cases already provide discrete symptom evidence, but TSA is not bypassed: its schema and validation path map released fields into the same typed present/absent/unknown predicates consumed by COTC. Thus DDXPlus exercises typed evidence normalization while not exercising slope or change-point estimation. On MIMIC-IV, TSA consumes irregular chemistry series and emits the discrete abnormal and trend predicates used by COTC, including the deterministic lab bins and available longitudinal summaries. The reported MIMIC ranking and completion results therefore use TSA-derived evidence tokens; they do not fit a cohort random-intercept model at test time. PhysioNet-style checks of which statistics are emitted appear in Appendix G.

3.3 COTC Module: Ternary Energies and Learning

The COTC module is the trainable core of COTCAgent. We do not train a language-model reader. We do not train the question policy. We train only the knowledge-base edge strengths ϕ and the absence scale γ that define the ternary energy below, under listwise ranking and a sparse curriculum.

Knowledge Base. We use a symptom, trend, and disease incidence matrix $M \in \{0,1\}^{|D|\times|E|}$, distinct from the TSA adapter f_{TSA} in Eq. 1. On DDXPlus we instantiate D and E from the public release with 49 pathologies and 223 evidences. On MIMIC-IV we build a co-occurrence knowledge base between phenotype groups and lab-derived evidence tokens on the training split only. Edges that never appear in training remain at their initialization and do not receive gradients.

Inverse Disease Frequency as a Fixed Sufficient Statistic. Common findings appear in many diseases and are weakly discriminative. Rare findings shrink the active set faster. For evidence e_j we define an IDF weight that depends only on knowledge-base incidence,

$$w_j = \log \frac{|D| + 1}{|\{d_i : M_{ij} = 1\}| + 1}. \tag{4}$$

The outer $\log w_j$ in Eq. 5 is extra flattening, not a second copy of IDF. The weight w_j already compresses incidence, and $\log w_j$ further flattens common findings. We clamp $w_j \geq 1$ before that outer log so that any finding present in more than about $1/e$ of diseases contributes 0 from the IDF term ($\log 1 = 0$) and is carried only by $\log \phi_{ij}$. This clamp is strong on dense co-occurrence bases such as MIMIC phenotype chapters. Without the clamp, $\log w_j < 0$ would invert ranking toward empty rare diseases. When population prevalences are unavailable, w_j acts as a surrogate log-odds weight that can be checked against a knowledge-base row and does not need a generative symptom model. Holding w_j fixed while learning ϕ separates two roles. IDF encodes how rare a finding is in the ontology. The matrix ϕ encodes how strongly that finding should support a particular disease after seeing labeled differentials.

Ternary Evidence and Additive Gibbs Energy. Let $P, A \in \{0,1\}^{|E|}$ be masks for present and absent evidences, and define unknown slots by $U = \mathbf{1} - P - A$, with present preferred if both flags are set.

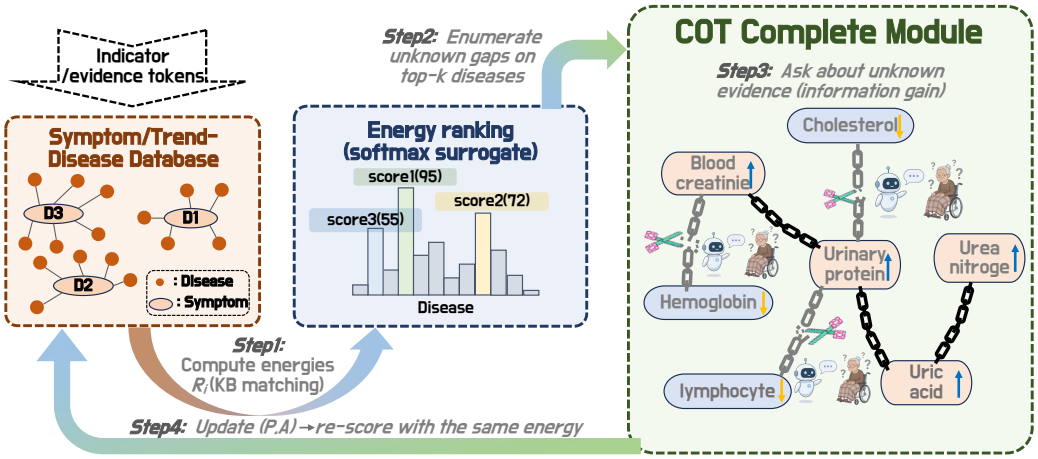


Figure 2: COTC scoring and completion. Structured cues feed the knowledge-base matcher. Ranked diseases, explicit unknown findings, and follow-up turns share the same energy in Algorithm 1.

Only known facts update the energy:

$$R_i = \sum_{j:M_{ij}=1} P_j \big(\log w_j + \log \phi_{ij} \big) - \sum_{j:M_{ij}=1} A_j \, \mathrm{softplus}(\gamma w_j), \tag{5}$$

$$\tilde{P}(d_i \mid P, A) = \frac{\exp(R_i/\tau)}{\sum_\ell \exp(R_\ell/\tau)}. \tag{6}$$

Here $\phi_{ij} = \mathrm{softplus}(\psi_{ij})$ is a learnable edge strength with free parameter ψ_{ij}, $\gamma > 0$ is a learnable absence scale, and τ is a temperature. Eq. 5 is an additive Gibbs energy over knowledge-base incidences. Positive findings contribute $\log w_j + \log \phi_{ij}$ on the log scale. Confirmed absences contribute a softplus penalty whose magnitude grows with IDF. Unknown evidences contribute exactly zero. That is the formal split between missingness and negation. Treating unknowns as absent would replace the zero contribution with a negative term for every unasked finding. Incomplete intakes would then collapse toward low-mass rankings, and completion would lose its value. The softmax in Eq. 6 is a ranking score, not a calibrated clinical posterior. An implicit uniform support prior can still bias toward rare diseases when patterns are ambiguous. That bias is acceptable for Hit@k but not for absolute risk reporting.

Training Objective, Gradients, and Curriculum. The optimized parameters are ψ, which determines $\phi = \mathrm{softplus}(\psi)$ on knowledge-base edges, and the scalar γ that scales soft absence penalties. We initialize $\phi_{ij} = \mathrm{softplus}(\psi_{ij})$ to 0.78 on incidence edges and γ to 0.15 by setting ψ and the unconstrained γ pre-image to $\log(\mathrm{expm1}(\cdot))$. Unused slots stay at that initialization because gradients are masked by M. The frozen-init rows in Table 1 use exactly this start and never update ψ or γ. IDF weights remain frozen functions of M. Given a minibatch of cases with gold pathology $d^\star$ and randomly masked views of the full evidence set, we minimize listwise cross-entropy

$$\mathcal{L}(\phi, \gamma) = -\sum_c \log \tilde{P}\big(d_c^\star \mid P_c, A_c\big) + \lambda \| \log \phi \|_2^2. \tag{7}$$

This loss is the negative log-likelihood of a Plackett-Luce style ranking model restricted to the top label. Gradients push mass onto $d^\star$ under the current ternary view. The ℓ_2 penalty on $\log \phi$ discourages arbitrarily sharp edges. Gradients on ϕ are masked to incidence edges with $M_{ij} = 1$, so unused slots stay at initialization and cannot absorb spurious correlations outside the knowledge base. A curriculum ramps the keep-fraction from 0.1 to 0.5 over training and, with probability 0.35, substitutes the released initial-evidence view, so that ϕ cannot spend the whole run on nearly complete charts (Appendix A). Supervision is ranking only. We never back-propagate through gap selection. The trained COTC module can therefore be reused across mass and information-gain policies. The corresponding sparse-curriculum training loop is Algorithm 2 in Appendix A.

Algorithm 1 Bounded probabilistic CoT completion in COTCAgent

Require: KB M, COTC parameters (ϕ, γ, τ), initial (P, A), threshold θ, max rounds $R_{\max}$, asks per round a, shortlist size m

```
Initialize evidence log L ← ∅
for r = 1 to R_max do
    Compute {R_i}, P̃, H via Eqs. 5 and 6
    if max_i P̃_i ≥ θ or no unknown gaps remain then
        return ranked diseases, L
    end if
    Shortlist top-m gaps by π_mass; select up to a by IG via Eq. 9
    Ask templated questions; parse answers into present or absent updates
    Append updates to L; refresh (P, A)
end for
return ranked diseases, L
```

3.4 Probabilistic CoT Completion

Given the current masks P and A, the COTC module computes $\tilde{P}$ and Shannon entropy $H = -\sum_i \tilde{P}_i \log \tilde{P}_i$. Candidate gaps are unknown evidences that are incident to the top-k diseases under $\tilde{P}$.

Mass-Weighted Priority. A simple baseline scores gap e_j by posterior mass on diseases that require it,

$$\pi_{\text{mass}}(e_j) = \sum_{i: M_{ij}=1} \tilde{P}_i \, \phi_{ij}, \tag{8}$$

restricted to slots with $U_j = 1$. This prefers questions that support currently likely diseases. It can be viewed as a first-order score that rewards edges already carrying mass. It does not explicitly maximize expected entropy reduction and can over-ask about findings that confirm the leader without separating near ties.

Information-Gain Surrogate for One-Step Acquisition. Let p_j be the current posterior mass on diseases incident to e_j, clipped to the interval from 0.05 to 0.95 for stability. After setting e_j present or absent in a hypothetical update we obtain entropies H_{yes} and H_{no} by recomputing Eqs. 5 and 6 under the two counterfactual masks. The surrogate entropy reduction is

$$\text{IG}(e_j) = H - \big(p_j H_{\text{yes}} + (1 - p_j) H_{\text{no}}\big). \tag{9}$$

Eq. 9 is a one-step acquisition heuristic under the ranking scores $\tilde{P}$. It is not Bayesian mutual information because p_j is not a calibrated symptom likelihood. The agent asks the unknown whose two counterfactual updates have the largest weighted entropy reduction under the same energy used for final ranking. Using p_j as the Bernoulli weight is an explicit approximation when a generative likelihood for answering yes is unavailable. Clipping avoids degenerate extremes when the shortlist is already peaked. We retain the short label COTC-IG for this surrogate throughout the experiments. In practice we evaluate Eq. 9 over a shortlist of mass-priority candidates, namely the top-m gaps under Eq. 8. A counterfactual rescore of one gap costs $O(|D|)$, so the whole shortlist costs $O(m\,|D|)$ in total rather than a sweep over every unknown evidence. Because the question policy is deterministic given (ϕ, γ), gains from training and gains from information-gain selection can be ablated separately.

Stopping and Oracles. Stopping uses the mass threshold θ, an empty gap set, or the round limit $R_{\max}$. Questions are templates over evidence names. Answers map to present or absent labels that are written into the evidence log $\mathcal{L}$. The threshold θ trades early stopping against leftover uncertainty. Lower values save questions but risk locking onto a peaked yet incorrect leader. Values near 0.85 sit at the empirical knee of Hit@1 versus cost on DDXPlus cold-start. An oracle that asks only among gold-positive evidences upper-bounds policy quality under perfect evidence selection. Algorithm 1 shares one COTC energy for ranking, gap scoring, and re-scoring after each answer. That shared energy is probabilistic CoT completion. All hyperparameters θ, τ, $R_{\max}$, a, and m are fixed at validation time and held constant across matched-budget comparisons unless a sweep is explicitly reported. Because answers enter $\mathcal{L}$ as discrete tokens, every ranked list can be replayed from the

Table 1: DDXPlus static ranking (%) under different evidence completeness (n=8000).

Model	Evidence	Hit@1	Hit@2	Hit@3	Hit@5
Learned ϕ	Full	98.8	99.5	99.9	100.0
Learned ϕ	Partial	90.0	95.6	97.4	98.8
Learned ϕ	Initial	36.3	50.6	59.6	72.5
Frozen init ϕ	Full	71.9	80.7	83.1	86.0
Frozen init ϕ	Partial	63.4	72.4	76.3	80.1
Frozen init ϕ	Initial	25.5	35.2	40.9	52.8

knowledge-base energy without calling a language model again. That is what makes the question budget and the ranked list jointly checkable.

4 Experiments

4.1 Experimental Settings

DDXPlus Tchango et al. (2022) is the primary benchmark. It provides structured differentials with initial evidence, full evidence sets, and gold pathology. We train ϕ and γ on up to 200k rows with the sparse curriculum in Algorithm 2, report static Hit@k on 8k held-out cases, and run consultation on 1k cold-start cases from initial evidence only. Unless noted otherwise, consultation uses five rounds and two questions per round with early-stop mass $\theta = 0.85$. Every policy shares this checkpoint, the gold present-or-absent oracle, and the stopping rules. MIMIC-IV v3.1 Johnson et al. (2023) is the secondary check. Admission labs become discrete tokens and multi-label phenotype groups, the split is patient-disjoint, and the knowledge base is counted on the training split only. Hit@1 asks whether the top phenotype lies in the gold set. This is a completion check under lab sparsity, not a bedside diagnosis benchmark. The 8k static initial-evidence Hit@1 of 36.3% and the 1k no-question Hit@1 of 37.9% come from different slices. Hyperparameters are in Appendix A. Baselines. Internal policies are no questions, random unknowns, frequency and IDF, mass-weighted gaps, information-gain gaps, and a gold-evidence oracle. Under the same budget we reimplement a BASD-style MLP Tchango et al. (2022); Xu et al. (2019), an AARLC-style entropy stopper Yuan & Yu (2024), a Diaformer-style Transformer Chen et al. (2022), fixed-KB information gain Guan & Baral (2021), and Qwen2.5-Coder 7B as a local ask-and-rank model Liu et al. (2024); Nori et al. (2025). These are protocol-matched copies, not official checkpoints. COTC and the learned askers share the 200k-row pool, the train/validation split, and 14 epochs. The 7B model is used because it emits parseable evidence IDs (Appendix I). We report Hit@1 and average questions over seeds 0, 1, and 2. Deterministic policies on the 1k slice have seed standard deviation 0. Random questioning is $56.3\% \pm 0.4$ Hit@1, and the gold-evidence oracle is $98.4\% \pm 0.3$.

4.2 Comparison Results

With the learned COTC parameters, full evidence yields 98.8% Hit@1, 99.5% Hit@2, and 99.9% Hit@3 on 8k DDXPlus cases (Table 1). Partial present-only evidence yields 90.0% Hit@1. Initial evidence alone yields 36.3% Hit@1. Freezing ϕ at initialization drops full Hit@1 to 71.9%, so training the COTC edge matrix is necessary for the ranking layer. The gap between full evidence and initial evidence, 98.8% versus 36.3%, is the room that completion must recover under cold start.

Table 2 and Figure 3 summarize cold-start consultation under the matched budget. Information gain reaches 88.1% Hit@1 with 3.61 questions. It beats mass at 84.0% with 3.86 questions, random at 56.3% with 5.45 questions, frequency at 56.1%, and IDF at 48.0%, while asking fewer questions than the non-adaptive baselines. An oracle that samples only gold evidences reaches 98.4% with 2.18 questions. Remaining room therefore sits in gap selection rather than in the trained COTC energy alone. Under perfect evidence choice the same R_i nearly recovers full-evidence static accuracy.

Table 3 places COTC against protocol-matched copies of recent acquisition methods on one shared cold-start slice of n=1000. BASD-MLP and AARLC-Ent should be read against their own intake ranking of 26.2% Hit@1, not against COTC's no-question score of 37.9%. Asking raises that dual-head classifier only to 31.1% (BASD) or 28.0% (AARLC, entropy stop at 1.85 questions). The gain

Table 2: DDXPlus cold-start consultation under a matched budget with 5 rounds and 2 questions per round (n=1000). Hit@1 (%) and average questions.

Policy	Hit@1	Avg. questions
None (no questions)	37.9	0.00
Random	56.3	5.45
Frequency	56.1	5.49
IDF	48.0	5.78
COTC-mass	84.0	3.86
COTC-IG	88.1	3.61
Oracle gold evidences	98.4	2.18

Table 3: Protocol-matched external baselines on DDXPlus cold-start (5 rounds, 2 questions per round). BASD, AARLC, Diaformer, BED, and COTC rows use n=1000. LLM-Ask uses Qwen2.5-Coder 7B on a seed-0 shuffle of 100 cases from the same test pool. On that 100-case subset COTC-IG reaches 87.0% Hit@1 (4.22 questions) versus LLM-Ask 14.0% (6.59 questions).

Method	Hit@1 (%)	Avg. questions
BASD-MLP	31.1	4.73
AARLC-Ent	28.0	1.85
Diaformer-Lite	41.6	4.00
BED-IG (fixed KB)	60.4	5.65
LLM-Ask (Qwen2.5-Coder 7B)	14.0	6.59
COTC-mass	81.2	4.12
COTC-IG (Ours)	86.2	3.87

is small, and neither matches the learned COTC energy. Diaformer-Lite improves modestly. Fixed-KB BED-IG reaches 60.4% Hit@1 while asking more than COTC. A local LLM that both asks and ranks under the same gold answer oracle is weakest at 14.0% Hit@1 on n=100 (seed-0 shuffle of the first 100 cases of that slice). On those same 100 cases, no-question COTC reaches 32.0%, COTC-mass 78.0% with 4.60 questions, and COTC-IG 87.0% with 4.22 questions. The LLM gap is therefore not an artifact of comparing n=100 against n=1000. Table 3 uses one shared slice of 1000 cases so that every external method sees the same patients. On that slice COTC-mass reaches 81.2% Hit@1 with 4.12 questions and COTC-IG reaches 86.2% with 3.87 questions, close to 84.0% and 88.1% on the main cold-start slice in Table 2. The gap is between two slices of the same pool, not a second consultation protocol.

Figure 3(a) shows that information gain pulls ahead within the first few rounds and keeps a gap over mass thereafter. Random and frequency baselines plateau far below even though they spend more questions. Figure 3(c) shows that the same policies collapse without learned ϕ. Completion cannot make up for an untrained COTC module. On the 1097 test admissions from MIMIC-IV in Figure 3(b), no questions yield 72.7% Hit@1, information gain reaches 79.4% with 2.41 questions, mass reaches 79.2% with 2.34 questions, and random reaches 77.9% with 2.58 questions. Mass and information gain are nearly tied here under coarser phenotype labels and denser lab co-occurrence than in DDXPlus. We treat MIMIC as a secondary check under real EHR noise, not as a prospective diagnosis claim.

Training is necessary (Figure 3(c)); intake fraction, stopping, and shortlist results are reported with question costs in Appendix Figures 4, 5, and 6, Tables 4–7, and Appendix D.

4.3 Discussion

The results support three scoped conclusions. First, ternary unknown states preserve the value of asking by distinguishing unobserved findings from confirmed absences. Collapsing these states would prematurely penalize plausible diseases, distort the initial ranking, and reduce the usefulness of follow-up questions. Second, learned edge strengths and gap selection are complementary: the former determines how observed evidence changes disease scores, whereas the latter deter-

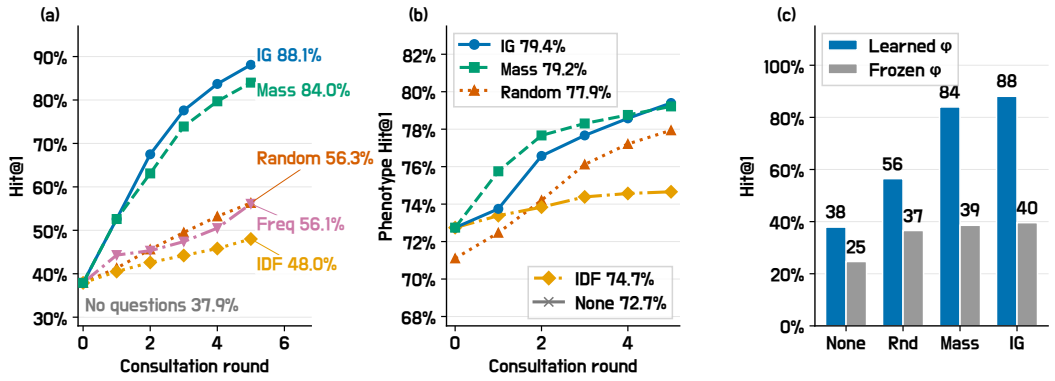


Figure 3: Main consultation results. (a) DDXPlus cold-start Hit@1 by round under a matched budget. (b) MIMIC-IV phenotype Hit@1 by round. (c) Learned versus frozen ϕ under matched DDXPlus policies. Bar heights are Hit@1.

mines which unresolved finding is most useful to acquire next. Neither component alone is sufficient—question selection cannot compensate for an inaccurate disease–evidence structure, while a well-trained scoring model cannot improve efficiently if it queries uninformative findings. Third, interactive methods must be compared under matched answer oracles, consultation rounds, question limits, and training-data exposure, because accuracy alone may conceal substantially different evidence costs or supervision advantages. The matched DDXPlus numbers make that split concrete. Information gain reaches 88.1% Hit@1 in 3.61 questions, while the gold-evidence oracle reaches 98.4% in 2.18 questions, so most of the remaining error is which unknown to ask. Frequency and IDF spend more questions, 5.49 and 5.78, yet stay near or below random at 56.1% and 48.0%. Prevalence and rarity therefore do not replace the learned scores. With frozen edges, the same information-gain rule reaches only 39.6% Hit@1 and uses 6.80 questions, close to the learned no-question score of 37.9%. AARLC stops earlier, at 1.85 questions and 28.0% Hit@1, so a smaller question count is not by itself a better consultation.

MIMIC-IV remains a secondary phenotype-ranking robustness check rather than a prospective diagnostic evaluation. TSA actively converts its irregular chemistry series into abnormality and trend predicates used by COTC. However, the coarser phenotype labels, denser laboratory co-occurrence, and retrospective observation process reduce the separation between mass- and information-gain policies. Their near tie should therefore be interpreted as evidence that policy advantages depend on label granularity and evidence structure, not as evidence that TSA is inactive or unnecessary. Appendix J provides the complete interpretation.

The method nevertheless assumes that answers can be mapped reliably to present or absent states, produces ranking scores rather than calibrated posterior probabilities, and remains an in-knowledge-base model. Performance on held-out diseases falls near random (Appendix E), indicating that information-gain questioning cannot recover disease–evidence relations that were never learned. The automatic answer oracle also removes linguistic ambiguity, patient refusal, measurement error, and real consultation costs. Consequently, the reported gains isolate the structured completion mechanism under controlled conditions; they do not establish open-world generalization, calibrated clinical risk, patient safety, or clinical utility.

5 Conclusion

COTCAgent casts preventive consultation as bounded probabilistic completion of a ternary evidence log under a public question budget. Separating TSA, a learnable knowledge-base energy, and deterministic gap selection keeps rankings tied to evidence that can be replayed rather than to free-form chain of thought. Training and information-gain completion are complementary. Structured completion can stay competitive without an LLM ranker under matched budgets. Calibrated scoring and prospective evaluation are left to future work.

ETHICS STATEMENT

DDXPlus is a synthetic, publicly released benchmark. MIMIC-IV is a de-identified clinical database accessed under its credentialing and data-use requirements. All required institutional ethics approvals, researcher credentialing, and data-use permissions were obtained before the data were accessed and analyzed. COTCAgent is not a medical device, does not provide calibrated disease risk, and must not be used to make clinical decisions.

REPRODUCIBILITY STATEMENT

The supplement specifies data splits, evidence construction, training schedules, seeds, stopping rules, and the shared answer oracle.

AI USE STATEMENT

Language models used as evaluated askers or readers are named in the experimental protocol and are part of the object of study. The authors also used writing assistants for implementation support, consistency checking, and language polishing. The authors verified the mathematical statements, experimental records, citations, and final manuscript and take responsibility for the submitted content.

REFERENCES

Junying Chen, Dongfang Li, Qingcai Chen, Wenxiu Zhou, and Xin Liu. Diaformer: Automatic diagnosis via symptoms sequence generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(4):4432–4440, 2022. doi: 10.1609/aaai.v36i4.20365.

Lauren Nicole DeLong, Ramon Fernández Mir, Zonglin Ji, Fiona Niamh Coulter Smith, and Jacques D. Fleuriot. Neurosymbolic AI for reasoning on biomedical knowledge graphs. *arXiv preprint arXiv:2307.08411*, 2023. ICML 2023 Workshop on Knowledge and Logical Reasoning in the Era of Data-driven Learning.

Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. A guide to deep learning in healthcare. *Nature Medicine*, 25(1):24–29, 2019. doi: 10.1038/s41591-018-0316-z.

Mohammad Feli, Iman Azimi, Pasi Liljeberg, and Amir M. Rahmani. An LLM-powered agent for physiological data analysis: A case study on PPG-based heart rate estimation. In *2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1–7, 2025. doi: 10.1109/EMBC58623.2025.11254428. URL https://doi.org/10.1109/EMBC58623.2025.11254428.

Álvaro García-Barragán, Ahmad Sakor, Maria-Esther Vidal, Ernestina Menasalvas, Juan Cristobal Sanchez Gonzalez, Mariano Provencio, and Víctor Robles. NSSC: a neuro-symbolic AI system for enhancing accuracy of named entity recognition and linking from oncologic clinical notes. *Medical & Biological Engineering & Computing*, 63(3):749–772, 2025. doi: 10.1007/s11517-024-03227-4.

Hong Guan and Chitta Baral. A bayesian approach for medical inquiry and disease inference in automated differential diagnosis. *arXiv preprint arXiv:2110.08393*, 2021.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023. doi: 10.1038/s41597-022-01899-x.

Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

Jiaheng Liu, Dawei Zhu, Zhiqi Bai, Yancheng He, Huanxuan Liao, Haoran Que, Zekun Wang, Chenchen Zhang, Ge Zhang, Jiebin Zhang, Yuanxing Zhang, Zhuo Chen, Hangyu Guo, Shilong Li, Ziqiang Liu, Yong Shan, Yifan Song, Jiayi Tian, Wenhao Wu, Zhejian Zhou, Ruijie Zhu, Junlan Feng, Yang Gao, Shizhu He, Zhoujun Li, Tianyu Liu, Fanyu Meng, Wenbo Su, Yingshui Tan, Zili Wang, Jian Yang, Wei Ye, Bo Zheng, Wangchunshu Zhou, Wenhao Huang, Sujian Li, and Zhaoxiang Zhang. A comprehensive survey on long context language modeling. *arXiv preprint arXiv:2503.17407*, 2025.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. AgentBench: Evaluating LLMs as agents. In *International Conference on Learning Representations*, 2024.

Haipeng Luo, Huawen Feng, Qingfeng Sun, Can Xu, Kai Zheng, Yufei Wang, Tao Yang, Han Hu, and Yansong Tang. AgentMath: Empowering mathematical reasoning for large language models via tool-augmented agent. *arXiv preprint arXiv:2512.20745*, 2025.

Harsha Nori, Mayank Daswani, Christopher Kelly, Scott Lundberg, Marco Tulio Ribeiro, Marc Wilson, Xiaoxuan Liu, Viknesh Sounderajah, Jonathan M. Carlson, Matthew P. Lungren, Bay Gross, Peter Hames, Mustafa Suleyman, Dominic King, and Eric Horvitz. Sequential diagnosis with language models. *arXiv preprint arXiv:2506.22405*, 2025.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. ChatDev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15174–15186, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.810.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. ToolLLM: Facilitating large language models to master 16000+ real-world APIs. In *International Conference on Learning Representations*, 2024.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, Juanma Zambrano Chaves, Szu-Yeu Hu, Mike Schaekermann, Aishwarya Kamath, Yong Cheng, David G. T. Barrett, Cathy Cheung, Basil Mustafa, Anil Palepu, Daniel McDuff, Le Hou, Tomer Golany, Luyang Liu, Jean-Baptiste Alayrac, Neil Houlsby, Nenad Tomasev, Jan Freyberg, Charles Lau, Jonas Kemp, Jeremy Lai, Shekoofeh Azizi, Kimberly Kanada, SiWai Man, Kavita Kulkarni, Ruoxi Sun, Siamak Shakeri, Luheng He, Ben Caine, Albert Webson, Natasha Latysheva, Melvin Johnson, Philip Mansfield, Jian Lu, Ehud Rivlin, Jesper Anderson, Bradley Green, Renee Wong, Jonathan Krause, Jonathon Shlens, Ewa Dominowska, S. M. Ali Eslami, Katherine Chou, Claire Cui, Oriol Vinyals, Koray Kavukcuoglu, James Manyika, Jeff Dean, Demis Hassabis, Yossi Matias, Dale Webster, Joelle Barral, Greg Corrado, Christopher Semturs, S. Sara Mahdavi, Juraj Gottweis, Alan Karthikesalingam, and Vivek Natarajan. Capabilities of Gemini models in medicine. *arXiv preprint arXiv:2404.18416*, 2024.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R. Pfohl, Heather Cole-Lewis, Darlene Neal, Qazi Mamunur Rashid, Mike Schaekermann, Amy Wang, Dev Dash, Jonathan H. Chen, Nigam H. Shah, Sami Lachgar, Philip Andrew Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Agüera y Arcas, Nenad Tomašev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle K. Barral, Dale R. Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950, 2025. doi: 10.1038/s41591-024-03423-7.

Arsène Fansi Tchango, Rishab Goel, Zhi Wen, Julien Martel, and Joumana Ghosn. DDXPlus: A new dataset for automatic medical diagnosis. In *Advances in Neural Information Processing Systems*, volume 35, pp. 31306–31318, 2022. doi: 10.52202/068431-2270. URL https://proceedings.neurips.cc/paper_files/paper/2022/

hash/cae73a974390c0edd95ae7aeae09139c-Abstract-Datasets_and_Benchmarks.html.

Zifeng Wang, Zichen Wang, Balasubramaniam Srinivasan, Vassilis N. Ioannidis, Huzefa Rangwala, and Rishita Anubhai. BioBridge: Bridging biomedical foundation models via knowledge graphs. In *International Conference on Learning Representations*, 2024.

Michael Wornow, Suhana Bedi, Miguel Angel Fuentes Hernandez, Ethan Steinberg, Jason Alan Fries, Christopher Ré, Sanmi Koyejo, and Nigam H. Shah. Context clues: Evaluating long context models for clinical prediction tasks on EHRs. *arXiv preprint arXiv:2412.16178*, 2024.

Lin Xu, Qixian Zhou, Ke Gong, Xiaodan Liang, Jianheng Tang, and Liang Lin. End-to-end knowledge-routed relational dialogue system for automatic diagnosis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7346–7353, 2019. doi: 10.1609/aaai.v33i01.33017346.

Ran Xu, Patrick Jiang, Linhao Luo, Cao Xiao, Adam Cross, Shirui Pan, Jimeng Sun, and Carl Yang. A survey on unifying large language models and knowledge graphs for biomedicine and healthcare. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 6195–6205, 2025. doi: 10.1145/3711896.3736556.

Hongyi Yuan and Sheng Yu. Efficient symptom inquiring and diagnosis via adaptive alignment of reinforcement learning and classification. *Artificial Intelligence in Medicine*, 148:102748, 2024. doi: 10.1016/j.artmed.2023.102748.

Shuang Zhou, Zidu Xu, Mian Zhang, Chunpu Xu, Yawen Guo, Zaifu Zhan, Yi Fang, Sirui Ding, Jiashuo Wang, Kaishuai Xu, Liqiao Xia, Jeremy Yeung, Daochen Zha, Dongming Cai, Genevieve B. Melton, Mingquan Lin, and Rui Zhang. Large language models for disease diagnosis: a scoping review. *npj Artificial Intelligence*, 1(1):9, 2025. doi: 10.1038/s44387-025-00011-z.

Appendix

A Additional experimental details

This appendix expands protocols that the main text summarizes only briefly, so that numbers, splits, and stopping rules can be reproduced without ambiguity.

Training. The reported DDXPlus checkpoint is the one trained with the cold-start curriculum. It is trained on 200k training rows for 14 epochs with learning rate 0.015 under a cosine schedule, batch size 512, and weight decay 10^{-4}. Sparse masking uses absent fraction 0.25 and initial-mix probability 0.35. When the curriculum is on, the keep fraction ramps from 0.1 to 0.5 across epochs rather than staying at its nominal value of 0.25; the absent fraction remains 0.25. With probability 0.35 a minibatch case is replaced by its released initial-evidence view, otherwise it is a randomly thinned view of the full evidence set. Optimization uses listwise cross-entropy on the gold pathology as in Eq. 7 of the main text. We select the checkpoint by a composite of cold-start information-gain Hit@1, question cost, and static partial or initial scores among parallel training configurations, as shown in Figure 9.

Algorithm 2 Sparse-curriculum training of the COTC parameters (ϕ, γ)

Require: Cases $\{(E_c^{\text{full}}, d_c^\star)\}$, KB M, epochs E, keep/absent schedules, initial-mix probability p_{mix}

Initialize ψ_{ij} so that $\phi_{ij} = 0.78$ on edges with $M_{ij} = 1$; $\gamma \leftarrow 0.15$
for $e = 1$ to E **do**
  Sample minibatch; draw keep fraction k_e and absent fraction a_e
  for each case c **do**
    if $\text{Unif}(0,1) < p_{\text{mix}}$ **then**
      Use the released initial-evidence view of c
    else
      Sample present mask P_c from E_c^{full} with keep k_e; set A_c on a fraction a_e of remaining known negatives
    end if
    Compute R_i and $\tilde{P}$ via Eqs. 5 and 6; accumulate Eq. 7
  end for
  Update ψ, γ by AdamW; set $\phi = \text{softplus}(\psi)$
end for
return (ϕ, γ) selected by the validation composite in Eq. 10

Consultation Protocol. Unless stated otherwise, cold-start cases begin with only the released initial evidence and treat every other finding as unknown rather than absent. Each round asks up to two unknown evidences. Answers follow the gold evidence set: a finding is labeled present if it appears for the case and absent otherwise. Early stopping uses mass threshold θ=0.85 when enabled, and the sweeps below vary that threshold while holding the ask and round budget fixed. All policies share the same learned checkpoint and the same answer oracle, so differences in Hit@1 and average questions are attributable to gap selection rather than to inconsistent answering.

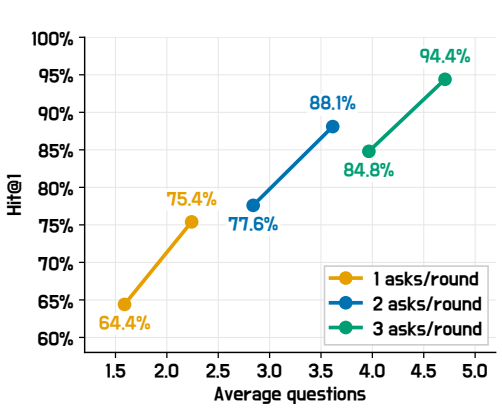
Why Static and Consultation Hit@1 Differ at Intake. Table 1 in the main text reports initial-evidence Hit@1 of 36.3% on an 8k static slice, whereas Table 2 reports no-question Hit@1 of 37.9% on a 1k cold-start consultation subset. These are not contradictory: they evaluate different case draws under the same trained energy. Throughout the paper we therefore treat static ranking as a completeness probe and consultation as the matched-budget interactive evaluation, and we avoid equating the two intake numbers.

MIMIC Construction. We sample 6000 admissions from MIMIC-IV v3.1 and keep a fixed panel of 18 common lab item IDs: creatinine, BUN, glucose, electrolytes, CBC, anion gap, albumin, bilirubin, lactate, troponin T, INR, and PTT. Values are binned into discrete abnormal or trend tokens. Eight phenotype groups are taken from diagnosis-code chapters: circulatory, endocrine, genitourinary, infectious, respiratory, digestive, neoplasm, and injury. Requiring at least three laboratory measurements and at least one extractable evidence token leaves 4552 admissions. We then retain

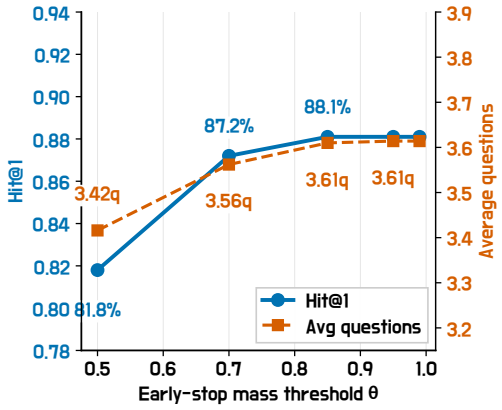
an admission only when its diagnosis codes include at least one of these eight chapters, which leaves 4384 admissions. These admissions are split by subject identifier so that every admission of one patient is assigned entirely to training or entirely to test. The split contains 3287 training admissions and 1097 test admissions. The knowledge base is phenotype and evidence co-occurrence counted on the training admissions. The reported metric asks whether the top-1 phenotype lies in the gold phenotype set for that admission, which is a completion check under sparse labs rather than a claim of primary diagnosis accuracy. The consultation numbers in the main text, no questions at 72.7% and information gain at 79.4% with 2.41 questions, are measured on the 1097 test admissions.

B BUDGET AND STOPPING SWEEPS

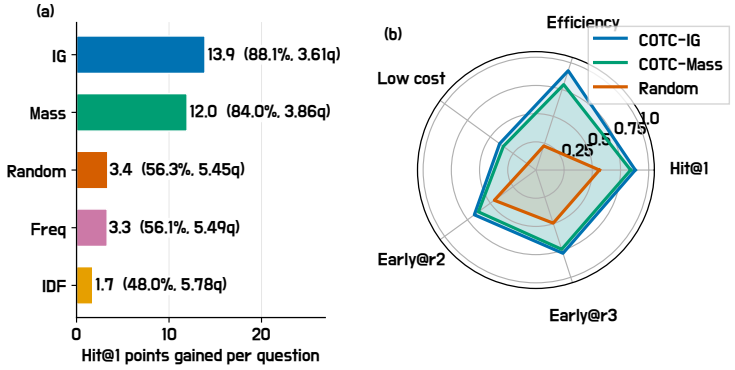
The main text reports one operating point with two asks and five rounds at θ=0.85. Figure 4 collects the asks/rounds Pareto, the θ sweep, and the efficiency view; numeric cells remain in Tables 4 and 5. Table 6 randomly drops initial evidences to stress sparse intakes: keep probability 1.0 recovers the main-text cold start, keep probability 0.5 removes half of the released intake at random, and keep probability 0.0 starts from a fully empty evidence log. Round-wise and learning results appear in main-text Figure 3. Figure 6 plots the keep-probability stress test, and Figure 5 shows curriculum training together with intake ablations.



(a) Asks/rounds budget Pareto



(b) Early-stop threshold θ



(c) Policy efficiency

Figure 4: Budget, stopping, and efficiency on DDXPlus cold-start. (a) Asks/rounds Pareto (Table 4). (b) Early-stop θ (Table 5). (c) Hit@1 points per question and multi-metric radar; information gain leads.

Reading Table 4 row by row clarifies the Pareto claim. Moving from one ask and three rounds to two asks and five rounds raises Hit@1 from 64.4% to 88.1% while average questions rise from 1.59 to 3.61. Adding a third ask under five rounds further raises Hit@1 to 94.4% at 4.71 questions, which is useful when a higher budget is available but is no longer the matched setting used for policy comparisons in Table 2. Table 5 shows that lowering θ to 0.50 stops earlier and loses accuracy,

whereas raising θ above 0.85 no longer changes Hit@1 or question cost on this checkpoint, which is why 0.85 is used as the default operating point. Figure 4(c) reports Hit@1 points gained per question and a multi-metric radar on the same matched budget; information gain leads on both views, consistent with selecting gaps that sharpen $\tilde{P}$ under Eq. 9.

Table 4: COTC-IG budget sweep on DDXPlus cold-start.

Asks/round	Rounds	Hit@1 (%)	Avg. questions
1	3	64.4	1.59
1	5	75.4	2.24
2	3	77.6	2.84
2	5	88.1	3.61
3	3	84.8	3.97
3	5	94.4	4.71

Table 5: Early-stop mass threshold θ with 2 asks and 5 rounds.

θ	Hit@1 (%)	Avg. questions
0.50	81.8	3.42
0.70	87.2	3.56
0.85	88.1	3.61
0.95	88.1	3.61
0.99	88.1	3.61

Table 6: Initial-evidence keep probability under COTC-IG.

Keep prob.	Hit@1 (%)	Avg. questions
1.0	88.1	3.61
0.5	79.3	5.43
0.0	68.9	7.41

C LEARNED VERSUS FROZEN CONSULTATION

The main text argues that completion and learning interact. Table 7 makes that claim quantitative by comparing the learned checkpoint to a frozen-ϕ initialization under the same policies, with the bar view in main-text Figure 3(c). Without training, information gain only reaches 39.6% Hit@1 and spends more questions than the learned run, which shows that gap selection cannot recover from an untrained energy. The learned rows match the main matched-budget table except for the random policy. This single-run ablation records 56.5% Hit@1 and 5.44 questions, against the multiseed mean of 56.3% and 5.45 questions in the main text. Mass and information-gain rows match exactly across the two tables. Figure 7 expands static Hit@k for both models as heatmaps so that the ranking-layer gap is visible before any questioning begins.

D LLM MATCHED-BUDGET DETAILS

The main text compares a local LLM that both asks and ranks on the full disease list (Table 3: n=100, five rounds, two asks per round). This section reports a separate shortlist protocol, so the two comparisons are not read as the same experiment. Table 9 uses n=60, at most three rounds, two asks per round, and a shortlist of 15 diseases. The language model is Qwen2.5-Coder 7B served locally through Ollama, chosen because the protocol requires parseable evidence IDs rather than fluent counseling text. All arms share the gold yes-or-no answer oracle used by COTC, so the language model never invents present or absent labels that COTC cannot also receive. Ranking and question-choice prompts are fixed templates. The ranker prompt asks for a single numbered disease

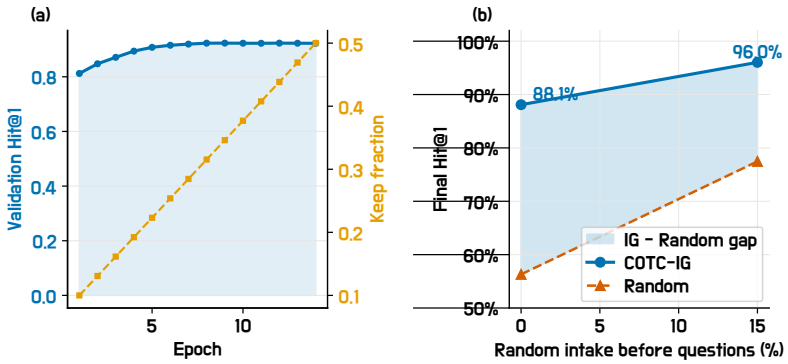


Figure 5: Training dynamics and intake ablation. (a) Cold-start curriculum: validation Hit@1 and keep fraction over epochs. (b) Final Hit@1 versus random intake fraction for COTC-IG and random questioning.

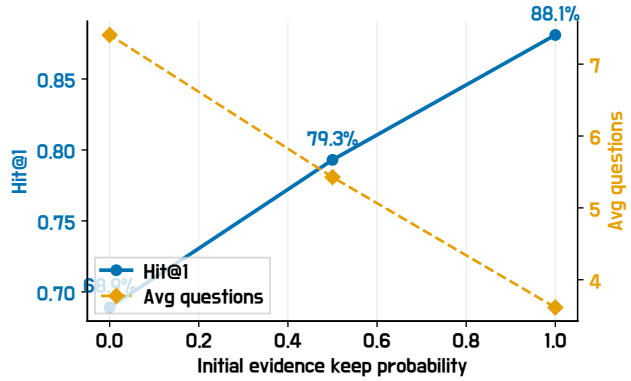


Figure 6: Initial-evidence keep probability stress test under COTC-IG, with Hit@1 on the left axis and average questions on the right axis.

from a candidate list given present and absent findings and forbids explanation. The asker prompt lists at most 40 unknown evidences with IDs and English questions and requests a item numbers or evidence identifiers. The templates used at evaluation time are given in Table 8. Ellipses and bracketed fields are filled per case.

COTC-IG ranks with the structured energy only and never queries the language model for ranking. LLM-full-static ranks among all diseases from the intake text without questions. LLM-shortlist-static ranks the COTC top-15 without questions, which is a fairer reader baseline because the candidate set is already narrowed. LLM-shortlist-ask lets the language model choose questions and then rank the shortlist. An earlier hybrid let the language model rank a shortlist after COTC-IG questions, but it stopped when the selected label matched the gold label. That rule uses the test label, so the arm is excluded from the comparison and from all claims. The retained COTC-IG and LLM-shortlist arms stop without access to the test label, using their stated validation-fixed rule or the round limit.

Table 9 lists the corresponding numbers. COTC-IG reaches 73.3% Hit@1 with 3.13 questions. LLM-shortlist-ask reaches 46.7% with 2.37 questions, and full-list static ranking reaches 6.7%. These results support the architectural claim that, when answers are shared, the bottleneck is gap selection and energy scoring rather than free-form generation. They do not claim that larger proprietary models would fail under a different protocol.

Table 7: Cold-start consultation: learned ϕ versus frozen initialization.

Model	Policy	Hit@1 (%)	Avg. q
Learned	None (no questions)	37.9	0.00
Learned	Random	56.5	5.44
Learned	COTC-mass	84.0	3.86
Learned	COTC-IG	88.1	3.61
Frozen	None (no questions)	24.7	0.00
Frozen	Random	36.6	6.98
Frozen	COTC-mass	38.6	6.88
Frozen	COTC-IG	39.6	6.80

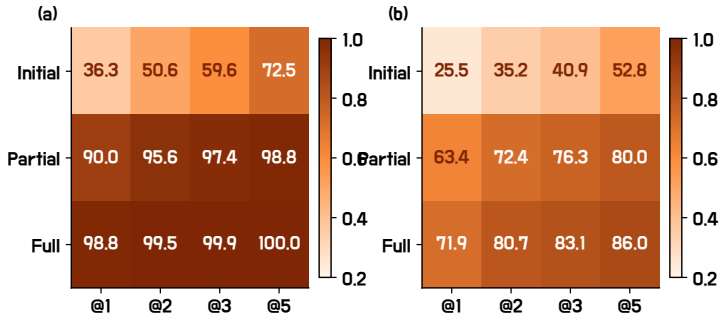


Figure 7: Static Hit@k heatmaps for learned ϕ in panel (a) and frozen initialization in panel (b) across evidence completeness.

E DISEASE HOLD-OUT

Figure 8 reports the disease hold-out stress test. Pathologies withheld from ϕ training remain near random under information-gain completion, while seen diseases retain strong gains. This pattern is expected for an in-knowledge-base completion model: edges that never receive ranking supervision cannot be recovered by questioning alone. We therefore treat hold-out as a negative control that bounds the claim, not as a failure of the matched-budget result on seen diseases. The main text does not claim open-world diagnosis, and this appendix is where that limit is made quantitative.

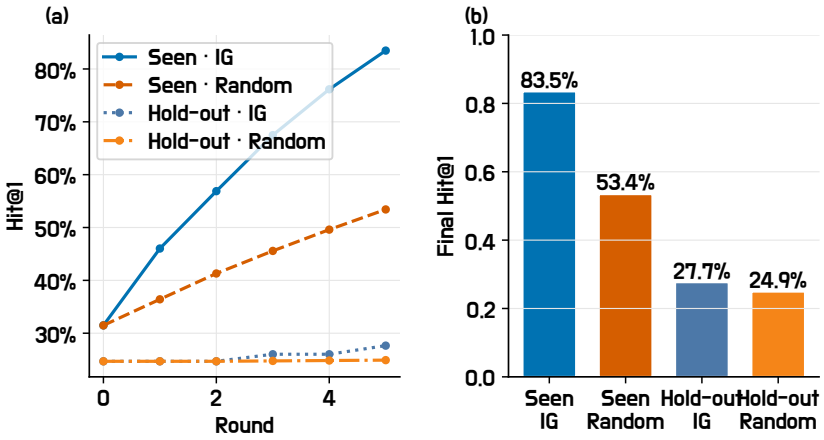


Figure 8: Disease hold-out. (a) Round-wise Hit@1 for seen versus hold-out subsets under information gain and random questioning. (b) Final Hit@1 summary.

Table 8: Fixed ranker and asker templates used at evaluation time.

Role	Template
Ranker	You are a careful clinical differential-diagnosis assistant. Given findings, pick the single most likely disease from the numbered list. Output format (strict): a single line like '3. DiseaseName'. Do not explain. Present findings: - ... Absent findings: - ... Candidates: 1. ... Answer:
Asker	Choose the most informative follow-up questions to refine the differential. Current findings: ... Top hypotheses: ... Pick $< a >$ item numbers from the list below. Reply with numbers or E_xxx ids only. 1. E_...: <English question> Picks:

Table 9: Label-free LLM comparison on the separate shortlist protocol (n=60). The hybrid that stopped on the gold label is excluded.

Policy	Hit@1 (%)	Avg. questions
COTC-IG	73.3	3.13
LLM shortlist ask	46.7	2.37
LLM shortlist static	23.3	0.00
LLM full static	6.7	0.00

F Training configuration selection

Figure 9 compares parallel training configurations with normalized metric profiles and selection scores. The cold-start curriculum checkpoint ranks first on the composite used for the main result. The unnormalized selection score is

$$s = \mathrm{Hit@1}_{\mathrm{IG}} - 0.01 \cdot \bar{q}_{\mathrm{IG}} + 0.05 \cdot \mathrm{Hit@1}_{\mathrm{partial}} + 0.05 \cdot \mathrm{Hit@1}_{\mathrm{initial}}, \tag{10}$$

which balances cold-start information-gain Hit@1 against question cost and static partial or initial accuracy. Other configurations may look stronger on a single axis, for example static full evidence alone, yet lose on cold-start consultation, which is the primary evaluation setting. The figure is therefore a selection audit rather than an exhaustive hyperparameter search.

G PhysioNet TSA grounding

DDXPlus does not exercise slope or change-point estimation, because its evidence is already discrete; TSA still maps those fields into typed predicates. To show that the executable-statistics layer is not vacuous when longitudinal inputs exist, Figure 10 summarizes TSA estimand composition on PhysioNet 2012 ICU series and a contingency heatmap of cohort-z anomaly flags versus in-hospital mortality. Panel (a) shows that the adapter emits a mixture of slope, change-point, and cohort-z tokens rather than one repeated token type; residual RMSE is attached as a STABLE/UNSTABLE quality flag on the slope token, and cohort-z values with $|z| > 2.5$ are the anomaly flags in panel (b). Panel (b) checks that anomaly flags co-occur with mortality in a structured way, which is a grounding sanity check rather than a predictive claim about PhysioNet itself. This appendix complements the two TSA paths used in the main experiments. DDXPlus passes through TSA's typed schema and validation path while omitting temporal estimators because its evidence is already discrete. MIMIC passes irregular laboratory series through TSA to produce the abnormal and trend predicates consumed by COTC. The PhysioNet analysis separately checks the slope, change-point, and cohort-z

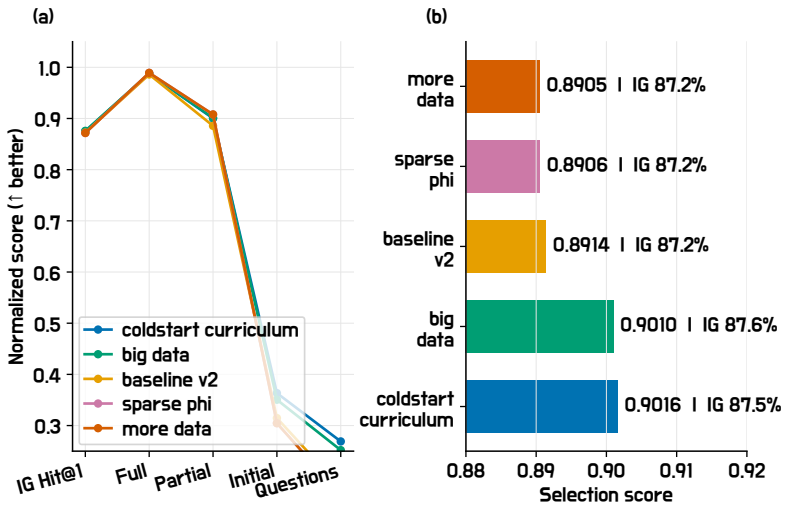


Figure 9: Configuration selection. (a) Parallel-coordinate style profiles with questions inverted so that higher is better. (b) Selection-score ranking with information-gain Hit@1 annotations.

emitters on another longitudinal source; it is a grounding audit rather than an additional diagnosis result.

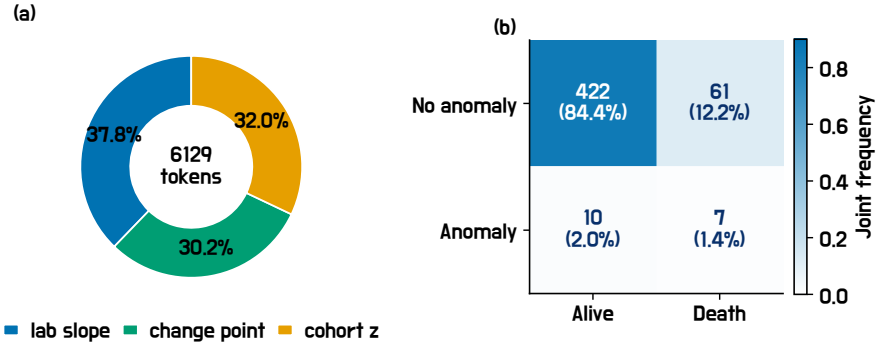


Figure 10: PhysioNet 2012 TSA outputs. (a) Donut of estimand token counts. (b) Joint frequency of anomaly flags and death.

H What we do not claim

We do not claim open-world diagnosis for held-out pathologies; hold-out Hit@1 stays near random because ϕ is never trained on those diseases, as shown in Figure 8. We do not claim that softmax scores are calibrated clinical probabilities. We do not use HealthBench or synthetic vignette leaderboards as primary evidence. MIMIC phenotype Hit@1 is a secondary completion check, not a substitute for prospective clinical evaluation. We also do not claim that the local 7B LLM comparison exhausts the space of language-model readers; it only shows that, under a shared answer oracle and matched ask budget, structured completion does not require an LLM ranker to remain competitive.

I External baseline reimplementations

Table 3 uses protocol-matched reimplementations rather than official author checkpoints, because public interactive DDXPlus releases rarely expose ternary unknown states, matched ask budgets,

and gold yes/no answer oracles in one package. BASD-MLP trains a two-layer MLP with a disease head and a next-evidence head on masked ternary views, then asks the top unknown scores and stops by mass threshold. AARLC-Ent reuses that backbone but stops when ranking entropy falls below a validation threshold, capturing the entropy-alignment idea of AARLC without full RL rollouts. Diaformer-Lite replaces the MLP with a two-layer Transformer encoder over present/absent/unknown tokens and the same dual heads. BED-IG runs information gain on a frozen knowledge-base energy with $\phi \approx 1$, isolating experimental design without learned edges. LLM-Ask uses local Qwen2.5-Coder 7B to choose questions and rank diseases under the same gold answers. All learned neural baselines and COTC are trained on the same 200k-row DDXPlus training pool for 14 epochs, use the same train/validation partition and masking views, and are evaluated on the same 1k cold-start cases with five rounds and two asks. This matches labeled-data exposure and epoch count; the architectures, parameter counts, and objectives remain method-specific. LLM-Ask uses $n{=}100$ because each case requires multiple Ollama calls. On the identical 100-case shuffle, COTC-IG reaches 87.0% Hit@1 with 4.22 questions (COTC-mass 78.0% / 4.60; no questions 32.0%), which is the fair subset comparison against LLM-Ask at 14.0% / 6.59.

J Extended discussion

Without training, information gain alone stalls near 39.6% Hit@1 (Figure 3(c) and Table 7), which serves as a weak no-training baseline. Figure 5(a) shows the cold-start curriculum. Figure 5(b) is a different probe: a random fraction of the full evidence set is revealed before questioning, and at a 15% fraction COTC-IG reaches 96.0% Hit@1 with 1.61 questions. That probe is easier than cold start, because the revealed tokens are drawn from the complete chart rather than from the released initial evidence, so it does not compete with the cold-start result of 88.1% Hit@1 at 3.61 questions under two questions per round and five rounds. The threshold $\theta = 0.85$ lies at the knee of the accuracy–cost trade-off (Figure 4, Tables 4 and 5). On the separate 60-case shortlist protocol (Appendix D), which is not the $n{=}100$ full-list comparison in Table 3, COTC-IG attains 73.3% Hit@1 versus 46.7% for the matched ask-and-rank reader.

The ternary unknown state is not a bookkeeping choice. Treating unasked findings as absent scores incomplete intakes as all-negative and removes the value of asking. Keeping unknown slots at zero energy leaves room for completion and makes cold-start ranked lists recoverable under a public question budget. Training and completion are coupled, not substitutes. Frozen edge strengths leave information-gain policies near the no-question regime, while the gold-evidence oracle still leaves room after training. Both a ranking COTC energy and a gap selector are required. Table 3 makes the same point with outside methods: protocol-matched reimplementations and a local LLM that both asks and ranks stay below learned COTC-IG when answers, rounds, and questions per round are held fixed.

Matched budgets matter because accuracy alone is incomplete. Figure 4 shows that information gain improves Hit@1 while spending fewer questions than random and frequency policies, and that the early-stop threshold has a clear knee. Reporting Hit@1 without question cost would hide the operating point that makes interactive methods comparable.

Two secondary settings constrain the scope. On MIMIC-IV phenotype ranking, mass and information gain remain nearly tied under coarser labels and denser lab co-occurrence; we treat that corpus only as a robustness check under real missingness, not as a primary diagnosis benchmark. Under the shared answer oracle, a local LLM reader also fails to outperform structured completion on the shortlist protocol. This comparison is limited to the named model and protocol and does not establish a general advantage over language-model readers.

Answers must be parseable into present or absent labels. Softmax scores are ranking scores, not calibrated posterior probabilities. DDXPlus is structured and synthetic, and MIMIC uses phenotype groups rather than a single primary code. Disease hold-out collapses near random for this in-knowledge-base model (Appendix E); frozen or unseen edges cannot be rescued by information gain. We therefore claim no open-world generalization, clinical utility, or safety under automatic answer oracles.

K REPRESENTATIVE CONSULTATION CASES

This section presents eight mechanism-oriented traces selected to cover short and long successes, correction by negative answers, policy contrast, closely related diseases, and explicit failures. The cases are illustrative rather than an additional evaluation set. DDXPlus traces use the main checkpoint trained with the cold-start curriculum, initial evidence, the gold-answer oracle, and at most five rounds with two questions per round. An unasked finding remains unknown and contributes zero energy; it is not treated as absent. The MIMIC traces use publicly available, de-identified Demo admissions so that individual question sequences can be inspected. Their strict case-study outcome requires Top-1 to equal the principal phenotype chapter, whereas the main MIMIC metric counts Top-1 membership in the admission's phenotype set. Consequently, these traces must not be used to reconstruct or reinterpret the main 79.4% result.

Table 10: Overview of the eight representative consultation traces. "Set hit" refers only to the main-text MIMIC phenotype-set criterion; the stricter case-study criterion requires the principal chapter at Top-1.

ID	Source	Gold target	IG result	Mechanism illustrated
D1	DDXPlus	Atrial fibrillation	2 questions, hit	Two near-neighbor dizziness predicates with opposite answers separate AF from PSVT.
D2	DDXPlus	GERD	6 questions, hit	Five negative answers move the ranking away from upper-respiratory and pericarditis explanations.
D3	DDXPlus	Pulmonary neoplasm	8 questions, hit	A misleading hemoptysis intake; IG succeeds while mass and random remain at viral pharyngitis after 10 questions.
D4	DDXPlus	Pulmonary embolism	4 questions, hit	DVT history and negative cardiopulmonary alternatives separate PE from edema.
D5	DDXPlus	Unstable angina	10 questions, miss	Stable and unstable angina share coarse evidence slots; the budget is exhausted without a discriminating predicate.
M1	MIMIC Demo	Infectious	4 questions, principal hit	Negative cohort-anomaly answers plus low albumin move Top-1 from injury/neoplasm to infectious.
M2	MIMIC Demo	Digestive only	10 questions, miss	Ten negative answers cannot overcome weak chapter-specific edges in the co-occurrence knowledge base.
M3	MIMIC Demo	Neoplasm	4 questions, principal hit	Negative cardiac and hematologic alternatives lift neoplasm above the competing chapters.

K.1 DDXPLUS QUESTION–ANSWER TRACES

Table 11 records representative oracle exchanges. For trajectories longer than the space available in the table, the listed exchanges are the decision-changing questions rather than a claim that omitted rounds did not occur. Probabilities are the normalized COTC ranking scores at the indicated point, not calibrated disease risks.

The D2 and D3 traces also clarify why realized entropy need not decrease monotonically. IG chooses a question by its *expected* reduction across possible answers; the observed answer may first move the case into a different disease basin or temporarily flatten the ranking. This is not inconsistent with the selection objective. Likewise, a clinically common symptom can be answered absent in an individual case: in D4, negative dyspnea does not mean the model encodes PE as a disease without dyspnea; it means that DVT history and the competing-diagnosis exclusions dominate for this particular gold record.

K.2 MIMIC DEMO QUESTION–ANSWER TRACES

The MIMIC examples in Table 12 use TSA-derived laboratory predicates. A raw value can be clinically striking without crossing the cohort-z threshold used by the corresponding cohort-anomaly

Table 11: Representative DDXPlus consultation trajectories. P/A denote oracle answers *present/absent*.

ID	Intake and initial ranking	Representative questions and oracle answers	Update and interpretation
D1	Palpitations; PSVT 0.164 vs. AF 0.149	"Slightly dizzy or lightheaded?"—P. "Lightheaded as if about to faint?"—A.	After one round, AF rises to 0.593 and entropy falls from 2.96 to 1.23. Similar wording carries different severity information and separates the near neighbors.
D2	Pain; viral pharyngitis 0.062; near-uniform entropy 3.67	Among six queries: rhinorrhea—A; fever—A; worse when lying flat—P; history of pericarditis—A; inspiratory pain—A.	The ranking moves through edema and pericarditis before reaching GERD at 0.41. Five of six answers are negative: the useful signal is exclusion, not symptom accumulation.
D3	Hemoptysis; viral pharyngitis 0.175; pulmonary neoplasm fifth at 0.102	Cough—A; swelling—A; loss of consciousness—A; three-month weight loss—P, together with the remaining IG-selected differentiators.	PE briefly leads, but the later weight-loss evidence lifts pulmonary neoplasm to 0.61 after eight questions. Mass and random both remain at viral pharyngitis after ten questions, showing the cost of reinforcing an incorrect initial leader.
D4	Swelling; edema 0.201 vs. PE 0.170	Relevant swelling location—P; DVT history—P; dyspnea—A; weight gain—A.	The first pair leaves edema and PE close (0.52 vs. 0.45); the negative alternatives then lift PE to 0.64. IG values the question because either answer can separate the two modes.
D5	Pain; gold unstable angina	Upper-respiratory findings—A; edema findings—A; dyspnea—P; overweight—P; red eyes—A; vaginal-discharge template—A; remaining questions exhaust the 10-question budget.	Final Top-1 is stable angina (0.185); unstable angina is fifth (0.057). The trace exposes both a vocabulary limit (no dedicated rest-vs.-exertion predicate) and a template-validity problem. The failure is budget exhaustion, not overconfident early stopping.

predicate. Thus an oracle answer of absent means that the named typed predicate is false under the prespecified transformation; it does not mean that the raw laboratory value is normal in every clinical sense.

Table 12: Representative MIMIC-IV Demo trajectories under the strict principal-chapter case-study criterion. Admission identifiers are de-identified MIMIC-IV Demo admission IDs.

ID	Admission and intake	Question sequence and oracle answers	Update and interpretation
M1	24181354; principal infectious; intake: abnormal lactate (3.1→1.4)	Round 0: albumin cohort anomaly—A; creatinine cohort anomaly—A. Round 1: WBC cohort anomaly—A; albumin abnormality—P.	Top-1 moves injury 0.134 → neoplasm 0.278 → infectious 0.311. Creatinine 7.9 and WBC 21.6 are large raw values but do not cross this Demo cohort's $\|z\| > 2.5$ predicate threshold.
M2	22585261; digestive is the only gold chapter; intake: albumin abnormality	Ten selected predicates, including troponin and cohort blood-cell abnormalities, are all answered A.	Neoplasm reaches about 0.35 and remains Top-1. IG, mass, and random all miss, exposing a knowledge-base limit rather than an IG-specific failure: the digestive chapter lacks sufficiently distinctive positive edges.
M3	29462354; principal neoplasm; intake: albumin abnormality	Troponin abnormality—A; troponin cohort anomaly—A; BUN shift—A; hemoglobin cohort anomaly—A.	Neoplasm rises from fifth to about 0.35 after four negative answers, while mass remains at circulatory after ten questions. The trace shows how exclusion of a cardiac explanation can be more useful than accumulating additional positive findings.

These case studies expose two limitations hidden by aggregate phenotype-set Hit@1. First, a MIMIC admission can count as correct in the main metric even when Top-1 is a comorbid chapter rather than the principal chapter; the stricter traces reveal whether questioning actually restores the principal label. Second, completion cannot invent a missing disease–evidence relation. M2 remains wrong after ten valid negative answers because the learned graph provides no sufficiently strong digestive-specific route. The examples therefore support the aggregate results while also delimiting what the reported metric and knowledge base can establish.