Who Should Be Consulted? Targeted Expert Selection for Rare Disease Diagnosis

Yinghao Fu [*1,2] **Chao Yang** [*1] **Xinye Chen** [1] **Yuting Yan** [1] **Shuang Li** [1]

Abstract

Effective collaboration between human experts and AI systems holds great promise in enhancing complex decision-making, particularly in challenging domains like *rare disease diagnosis*. In traditional Multi-Disciplinary Team (MDT) settings, human experts from different specialties are pre-assigned to review and discuss patient cases collaboratively. However, such fixed team structures may suffer from cognitive anchoring, incomplete knowledge, or misaligned expertise, especially when facing atypical or rare clinical presentations. In this paper, we propose **S**equential **E**xpert **E**ngagement for **R**are diseases (SEER) that *dynamically selects targeted human experts to participate in collaborative decision-making*. Our approach leverages a rule-based AI system with broad, structured medical knowledge to identify critical diagnostic paths and propose complementary expert inputs. The AI system serves two key roles: (1) recommending plausible diagnostic hypotheses and logical rules based on structured knowledge; and (2) identifying which experts, if consulted, are most likely to resolve diagnostic uncertainties. This targeted expert selection process helps avoid cognitive biases like anchoring and expands the decision space by inviting diverse, high-value perspectives. Moreover, the system is self-evolving, continuously updates its rules and its understanding of each expert's expertise based on newly collected data and feedback. Experiments on both synthetic and real-world rare disease datasets demonstrate that our framework improves diagnostic accuracy, reduces expert workload, and enhances the overall robustness of human-AI collaboration.

[*]Equal contribution [1] School of Data Science, The Chinese University of Hong Kong (Shenzhen) [2]Department of Biostatistics, City University of Hong Kong. Correspondence to: Shuang Li <lishuang@cuhk.edu.cn>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

1. Introduction

Diagnosing *rare diseases* presents a unique challenge: symptoms are often ambiguous and overlap with common conditions, leading to *long diagnostic delays* (The Lancet, 2024; Venus et al., 2025). Patients frequently consult multiple specialists before reaching a correct diagnosis, a delay that arises not from a lack of expertise, but from consulting the wrong experts at the wrong time (Evans, 2018).

Traditional methods such as Multi-Disciplinary Team (MDT) consultations, use fixed groups of clinicians (Qian et al., 2023). However, this static group-based approach is often inefficient (Lamb et al., 2012). Not all experts in the MDT may be relevant for a given case, and some may anchor their judgments on more common explanations due to cognitive biases, potentially derailing the diagnosis. Given the *high ambiguity* and *rarity* of these conditions, the **selection of the right expert to consult** becomes central to improving diagnostic accuracy and efficiency (Winters et al., 2021; Baynam et al., 2024). If we can identify and consult the most relevant specialist at each step, starting with one and progressively resolving ambiguity through targeted follow-up, we can significantly accelerate and improve the diagnostic process (Kurvers et al., 2023).

Yet, the question of "**Who should be consulted?**" remains under-explored in both research and clinical practice. Expert selection is usually handled manually or procedurally rather than based on a data-driven understanding of each specialist's domain strengths (Finn et al., 2022). This is especially problematic in rare disease settings where expertise is unevenly distributed and difficult to characterize (Domaradzki & Walkowiak, 2019).

To address this, we propose **SEER**: **Sequential Expert Engagement for Rare diseases**, a logic-driven, probabilistic framework that dynamically selects the most informative experts at each step of the diagnostic process. Unlike general-purpose AI decision systems or rigid MDT assignments, our system uses a self-evolving, rule-based reasoning model to suggest the next most informative expert to consult. This model does not aim to replace human specialists but rather serves as a **knowledge-driven coordinator**, grounded in probabilistic rules, to orchestrate expert involvement dynamically.

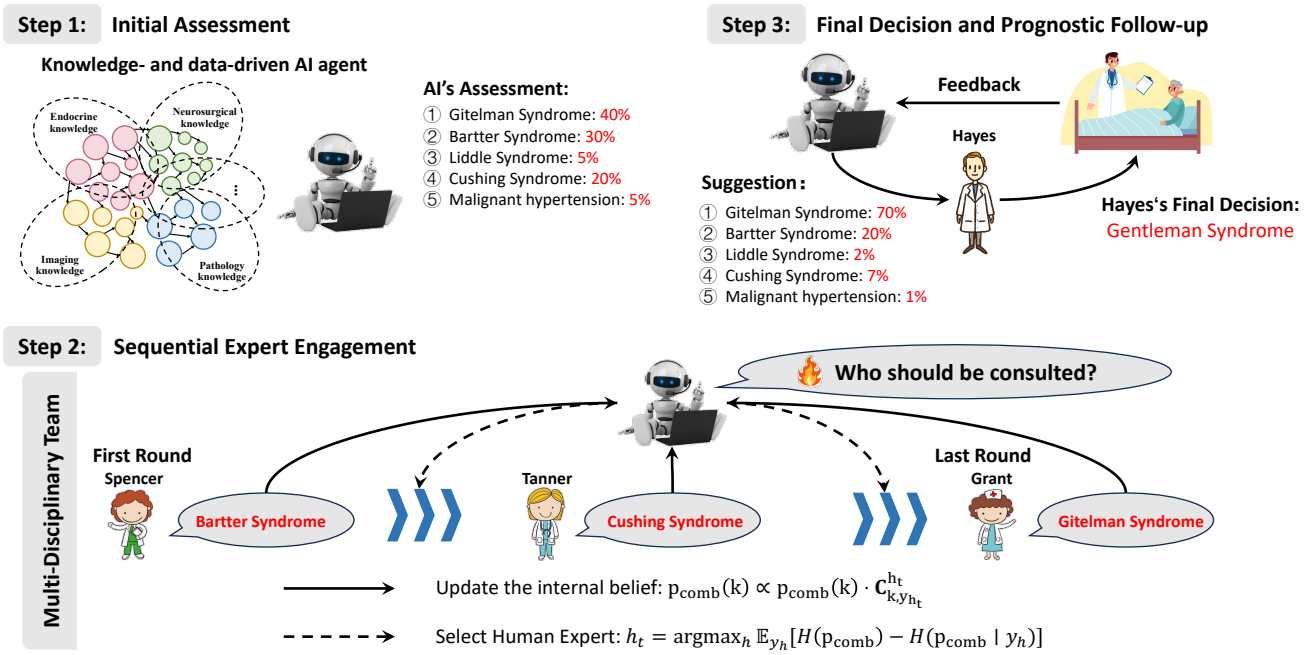

Figure 1. SEER Decision-Making Overview. For each new patient, SEER begins with an internal AI-generated prior over diagnoses. It then enters a closed-loop internal deliberation process and iteratively selects human experts based on the expected information gain, computed using confusion matrices, and refines its belief until uncertainty (entropy) falls below a threshold. No intermediate outputs are released; only the final calibrated diagnosis distribution is provided to clinicians.

Our SEER offers three core advantages:

(*i*) **Broad Knowledge Representation:** While not as specialized as a domain expert in any one field, our rule-based model captures a wide range of domain-specific logic, enabling it to recognize rare patterns and link them to appropriate specialties.

(*ii*) **Expertise Awareness:** Our model continuously learns a probabilistic representation of each human expert's strengths, allowing it to match cases to the most relevant clinicians.

(*iii*) **Self-Evolving Logic:** As more data and feedback accumulate, the system updates both its internal reasoning rules and its understanding of expert capabilities, making it adaptive to new diseases, changing team compositions, and evolving diagnostic protocols.

The primary objective of SEER is not to replace physicians in diagnosing rare diseases but to *minimize diagnostic delay and error by intelligently routing each patient through a minimally sufficient consultation pathway*. This addresses the core challenge in rare disease diagnosis, namely, finding the *right expert* at the *right time*, which is often overlooked in current decision-support systems. Just as assigning the wrong reviewer in academic peer review can lead to inconsistent or biased evaluations, assigning the wrong specialist can delay or misdirect clinical diagnosis, especially when symptoms overlap with more common diseases.

We also highlight that, when properly guided, the *anchoring effect*, typically viewed as a cognitive bias, *can serve a productive role*. If the *AI provides a high-quality probabilistic prior*, it can orient human experts toward more accurate decisions early in the diagnostic process, reducing ambiguity rather than introducing it.

Experiments on synthetic and real-world datasets, including a dataset related to the rare disease Gitelman syndrome, demonstrate that SEER outperforms MDT-based and random consultation strategies in both diagnostic speed and accuracy.

2. SEER: Sequential Expert Engagement for Rare diseases

SEER as a Confidence-Aware Coordinator SEER is not a co-decider but a *knowledge-driven*, *confidence-aware coordinator* that dynamically selects *which expert to query and when* to improve diagnostic accuracy. The overall algorithm is detailed in Algorithm 1.

For each patient case with features $\boldsymbol{x}$, SEER first forms an *initial probabilistic belief* over diagnosis labels $k \in [K]$, denoted as: $p^{\text{AI}}(k \mid \boldsymbol{x})$, which is based on historical data and logical rules (see Section 4 for details).

Human Modeling via Confusion Matrices In parallel, SEER maintains a personalized profile of each expert

Algorithm 1 SEER: Sequential Expert Engagement for Rare diseases

Inputs: Patient features $\boldsymbol{x}$, human experts' confusion matrices $\{\boldsymbol{C}^{(h)}\}_{h=1}^{N}$, threshold τ, max queries $H_{\max}$, calibration η
Initialize: Compute initial AI diagnosis distribution $p^{\mathrm{AI}}(k \mid \boldsymbol{x})$ for $k \in [K]$, set $p_{\mathrm{comb}} = p^{\mathrm{AI}}$
if $\max_{k\in[K]}\ p_{\mathrm{comb}}(k) \geq \tau$ **then**
  Output: $\arg\max_k p_{\mathrm{comb}}(k)$
  Terminate.
end if
for $t = 1$ to $H_{\max}$ **do**
  Select human expert h_t to maximize expected information gain:

$$h_t = \arg\max_h \mathbb{E}_{y_h}\left[H(p_{\mathrm{comb}}) - H(p_{\mathrm{comb}} \mid y_h)\right]$$

  Query the selected human expert h_t to obtain their opinion y_{h_t}
  Update p_{comb} using Bayes' rule:

$$p_{\mathrm{comb}}(k) \propto p_{\mathrm{comb}}(k) \cdot \boldsymbol{C}^{(h_t)}_{k,y_{h_t}}$$

  if $H(p_{\mathrm{comb}}) \leq \tau$ **then**
    Output: $\arg\max_k p_{\mathrm{comb}}(k)$
    Terminate.
  end if
end for
Output: $\arg\max_k p_{\mathrm{comb}}(k)$ after $H_{\max}$ iterations

$h \in [N]$, modeled via a confusion matrix $\boldsymbol{C}^{(h)} \in \mathbb{R}^{K\times K}$, defined as:

$$\boldsymbol{C}^{(h)}_{ij} = p(\text{Human } h \text{ says } j \mid \text{True label } i)$$

which characterizes the *conditional reliability* of the expert across diagnostic classes. Specifically,

- If human h is accurate across all domains, $\boldsymbol{C}^{(h)}$ is diagonal dominance: $\boldsymbol{C}^{(h)}_{k,k} \approx 1, \forall k \in [K]$.
- If the human specializes in a subset of labels $\mathcal{K}_s \subset [K]$, then the confusion matrix is diagonally dominant only within $\mathcal{K}_s$: $\boldsymbol{C}^{(h)}_{k,k} \approx 1, \forall k \in \mathcal{K}_s$, but $\sum_{j\neq k} \boldsymbol{C}^{(h)}_{k,j} \gg 0$ for $k \notin \mathcal{K}_s$.
- If the human systematically confuses between labels k and y, then $\boldsymbol{C}^{(h)}_{k,y} \gg 0,$ for some $k \neq y$.

Each confusion matrix $\boldsymbol{C}^{(h)}$ lets SEER identify expert h's strengths (reliable label subsets) and weaknesses (consistent misclassifications).

Confidence-Aware Expert Selection Set $p_{\mathrm{comb}} = p^{\mathrm{AI}}$. Before querying, SEER treats the expert response y_h as a random variable and selects the next expert by computing the *expected conditional entropy* of the predicted label distribution:

$$\mathbb{E}_{y_h}\left[H\left(p_{\mathrm{comb}} \mid y_h\right)\right] = -\sum_j p\left(y_h = j\right) \sum_k p\left(k \mid y_h = j\right) \log p\left(k \mid y_h = j\right)$$

where $p\left(y_h = j\right) = \sum_k p_{\mathrm{comb}}(k) \cdot \boldsymbol{C}^{(h)}_{kj}$ and $p\left(k \mid y_h = j\right) = \frac{p_{\mathrm{comb}}(k)\cdot \boldsymbol{C}^{(h)}_{kj}}{p(y_h=j)}$. SEER selects the expert h_t with the *lowest expected conditional entropy*, or equivalently, the one offering the *highest expected information gain*,

$$h_t = \arg\max_h H\left(p_{\mathrm{comb}}\right) - \mathbb{E}_{y_h}\left[H\left(p_{\mathrm{comb}} \mid y_h\right)\right].$$

Once an expert h is consulted, and a response y_h, such as j, is received, SEER updates its belief over the diagnoses using Bayes' rule:

$$p\left(k \mid y_h = j, \boldsymbol{x}\right) = \frac{p_{\mathrm{comb}}(k) \cdot \boldsymbol{C}^{(h)}_{kj}}{p\left(y_h = j \mid \boldsymbol{x}\right)}.$$

Theoretically, SEER's strategy resembles adaptive combinatorial expert selection, where each query reduces uncertainty in a submodular fashion. Despite only observing conditionally independent expert responses, such sequential routing can attain near-optimal diagnostic performance, as characterized by known results on submodular information-gathering (e.g., achieving a $(1 - 1/e)$-approximation to the best fixed expert combination) (Chen et al., 2015).

Query-Termination and Aggregation After each expert interaction, SEER updates its belief over possible labels. This iterative process continues until a stopping criterion (e.g., entropy threshold as shown in Algorithm 1) is met. Only then is a final decision made.

Importantly, the *internal belief* p_{comb} of SEER is **never revealed** to experts *during the interaction process*. The final aggregated diagnosis may be shown to the clinician **only after** expert querying is complete. This design reflects two considerations: 1) Revealing AI-generated beliefs or predictions to experts during the process *risks anchoring or influencing* their independent judgments, which corrupts the assumption of conditional independence necessary for accurate belief updating (Lee et al., 2015). 2) The question of *when and how to present AI recommendations to human experts still remains an open research problem* in the human-AI interaction area (Shaw et al., 2019). In many clinical contexts, deferring to the AI's opinion until after unbiased human assessments preserves the integrity of downstream reasoning (Yin et al., 2025; Wang et al., 2024).

3. Theoretical Analysis: Why Does SEER Lead to More Accurate Diagnosis?

The goal of SEER is not to replace physicians or directly provide diagnoses, but to **identify the right expert to consult at the right time**, routing each case through a *minimal yet sufficient sequence of expert consultations*. This reframes rare disease diagnosis as a problem of efficient expert selection under uncertainty, where the AI's broad but shallow prior is incrementally refined through targeted consultations. Our theoretical analysis formalizes this intuition by showing how *selective querying can systematically reinforce diagnostic accuracy*.

We will first show how SEER adapts its querying behavior to AI's current posterior belief $p_{\text{comb}}(y)$. As an illustration, let's first consider concrete examples to understand SEER's expert selection behavior:

Example 1 (Shallow but Broad Prior). In a 3-class problem ($K = 3$), suppose AI's prior is: $p^{\text{AI}}(A) = 0.4$, $p^{\text{AI}}(B) = 0.35$, and $p^{\text{AI}}(C) = 0.25$. Although weak, the AI believes A is slightly more likely. Suppose one human has $\boldsymbol{C}^{(1)}_{A,A} = 0.85$ and lower reliability for B and C. SEER queries this human to confirm or refute A, leveraging the weak signal in the AI's prior.

Example 2 (Ambiguity Resolution). Now suppose $p^{\text{AI}}(A) = 0.6$, $p^{\text{AI}}(B) = 0.3$, and $p^{\text{AI}}(C) = 0.1$. The main ambiguity lies between classes A and B. Suppose we have two human experts: 1) *Human 1 (A/B expert)*: $\boldsymbol{C}^{(1)}_{A,A} = 0.9$, $\boldsymbol{C}^{(1)}_{B,B} = 0.8$. 2) *Human 2 (B/C expert)*: $\boldsymbol{C}^{(2)}_{B,B} = 0.9$, $\boldsymbol{C}^{(2)}_{C,C} = 0.8$. In this case, SEER will correctly select Human 1 to resolve the uncertainty between A and B, accelerating convergence to the correct label.

Next, let's formally characterize SEER's expert selection pattern:

Theorem 3.1 (Characterization of SEER's Selection Behavior)**.** *Let $p^{(t)}_{\text{comb}}$ be the belief at iteration t. Let $\mathcal{H}$ be the pool of human experts with confusion matrices $\{\boldsymbol{C}^{(h)}\}_{h\in\mathcal{H}}$. Define the information gain from the consulting expert h as:*

$$\Delta H_h = H(p^{(t)}_{\text{comb}}) - \mathbb{E}_{Y_h\sim\boldsymbol{C}^{(h)}}[H(p^{(t+1)}_{\text{comb}}|Y_h)],$$

where $p^{(t+1)}_{\text{comb}}$ is updated by Bayes' rule. Then:

1. ***Reinforcement:*** *If $\exists y^*$ such that $p^{(t)}_{\text{comb}}(y^*) \geq 1-\epsilon$ ($\epsilon \ll 1$), and expert $h_{\text{reinforce}}$ satisfies $\boldsymbol{C}^{(h_{\text{reinforce}})}_{y^*,y^*} \geq 1-\delta$ ($\delta \ll 1$), then:*

$$h_{\text{reinforce}} = \arg\max_{h\in\mathcal{H}} \Delta H_h.$$

2. ***Disambiguation:*** *If $p^{(t)}_{\text{comb}}(y_1) \approx p^{(t)}_{\text{comb}}(y_2)$ for top candidates y_1, y_2, and expert $h_{\text{disambiguate}}$ satisfies $\boldsymbol{C}^{(h_{\text{disambiguate}})}_{y_1,y_1}, \boldsymbol{C}^{(h_{\text{disambiguate}})}_{y_2,y_2} \geq 1-\delta$, then:*

$$h_{\text{disambiguate}} = \arg\max_{h\in\mathcal{H}} \Delta H_h.$$

Detailed proof of the Theorem 3.1 can be found in the Appendix B. From the proof, we see that even when the Al provides a weak but informative prior over K classes at time t,

$$p^{(t)}_{\text{comb}}(y) = \begin{cases} \frac{1}{K} + \eta & \text{if } y = y^*, \\ \frac{1}{K} - \frac{\eta}{K-1} & \text{if } y \neq y^*, \end{cases}$$

where $\eta > 0$. This broad yet shallow prior, which slightly favors the true label y^*, acts as a valuable informational anchor that steers posterior inference toward y^* and reduces expected entropy after expert feedback with confusion matrix $C^{(h)}$. Hence, the expected information gain $\Delta H_h\left(p^{(t)}\right)$ surpasses that from a uniform prior. SEER leverages this by adaptively selecting experts who reinforce confident AI predictions via high diagonal accuracies or disambiguate close competitors through maximal confusion matrix discriminability. Thus, even with broad uncertainty, the Al's shallow but informative prior improves expert selection and collaborative inference beyond random guessing.

Next, we show that a good AI prior can accelerate the convergence to the true y^* and SEER will have a linear convergence rate.

Theorem 3.2 (Exponential Entropy Decay and Linear Convergence)**.** *Consider a classification problem with K classes and true label y^*. Assume the following:*

1. *The AI prior at round $t = 0$ satisfies a weak but informative bias, where $p^{\text{AI}}(y^*) \geq \frac{1}{K} + \eta$ for some $\eta > 0$, and the remaining probability mass is uniform over other classes, such that $p^{\text{AI}}(y) = \frac{1-p^{\text{AI}}(y^*)}{K-1}$ for all $y \neq y^*$.*

2. *There exists at least one human expert h whose confusion matrix $C^{(h)}$ satisfies local accuracy: $C^{(h)}_{y^*,y^*} \geq 1-\delta$, for small $\delta > 0$, and partial correctness: $C^{(h)}_{y^*,y^*} > C^{(h)}_{k,y^*}$, $\forall k \neq y^*$.*

3. *At each round t, SEER selects the expert h_t that maximizes expected entropy reduction: $h_t = \arg\max_h \Delta H_h$, where $\Delta H_h := H\left(p^{(t)}_{\text{comb}}\right) - \mathbb{E}[H(p^{(t+1)}_{\text{comb}})]$.*

Then:

1. *There exists a constant $\gamma > 0$ such that the expected posterior entropy decreases at least by γ each round:*

$$\mathbb{E}\left[H(p^{(t+1)}_{\text{comb}})\right] \leq H(p^{(t)}_{\text{comb}}) - \gamma.$$

2. *Defining the log-odds ratio for the true label:*

$$\mathcal{L}_t := \log \frac{p^{(t)}_{\text{comb}}(y^*)}{1 - p^{(t)}_{\text{comb}}(y^*)},$$

there exists a $\Delta_{\min} > 0$ *such that* $\mathbb{E}[\mathcal{L}_{t+1} \mid \mathcal{L}_t] \geq \mathcal{L}_t + \Delta_{\min}$. *Hence,* $\mathbb{E}[\mathcal{L}_t] \geq \mathcal{L}_0 + t\Delta_{\min}$.

3. *Consequently, after* T *rounds,*

$$\mathbb{E}[H(p^{(T)}_{\text{comb}})] \leq H(p^{(0)}_{\text{comb}}) - T\gamma,$$

guaranteeing linear convergence in entropy and exponential convergence of posterior confidence to 1. Moreover, the number of rounds to reach confidence level τ *satisfies:*

$$T \geq \frac{\log \frac{\tau}{1-\tau} - \mathcal{L}_0}{\Delta_{\min}},$$

where $\mathcal{L}_0$ *is the log-odds ratio for AI prior.*

The sketch of the proof can be found in Appendix. B.

4. How will SEER Self-Evolve Given Patient Feedback?

4.1. Rule Evolution On-the-Fly Given Patient Feedback

The effectiveness of SEER depends heavily on the quality of its rule-driven predictor p^{AI}. Initially, this predictor can be constructed from expert knowledge, providing a clear and interpretable foundation. Importantly, SEER can continuously **self-evolve** by reinforcing and updating its rules on the fly, using limited **reward-style feedback** $\tilde{r}_k \in \{-1, +1\}$. This feedback indicates only whether a prediction was correct (e.g., patient recovers) or incorrect (e.g., patient does not recover), without revealing the true label. We leverage these outcome-based signals to iteratively refine both the rules and their weights.

For each class $k \in [K]$, SEER maintains:

1. A rule set Γ_k, consisting of logical conditions that trigger a prediction for class k, for example:

$$\text{Label } k \leftarrow \mathbb{I}(\text{LabResult}(\boldsymbol{x})) \wedge \mathbb{I}(\text{GeneticPredisposition}(\boldsymbol{x}))$$

where $\mathbb{I}(\cdot)$ is the indicator function.

2. A weight vector $\boldsymbol{w}_k \in \mathbb{R}^d$ assigning importance to each rule.

Given input $\boldsymbol{x}$, a Boolean feature vector $\boldsymbol{\phi}_k(\boldsymbol{x}) \in \{0,1\}^d$ indicates which rules are activated. The binary predictor estimates the probability that label k is correct as $p_k(\boldsymbol{x}) = \sigma\left(\boldsymbol{w}_k^\top \boldsymbol{\phi}_k(\boldsymbol{x})\right)$, where $\sigma(z) = 1/(1+e^{-z})$ is the sigmoid function.

Loss Function for Rule Weight Learning We use a logistic loss defined over the binary feedback $\tilde{r}_{k,n} \in \{-1, +1\}$, where $+1$ indicates a correct prediction and -1 an incorrect one, for sample n:

$$\ell_k(\boldsymbol{w}_k) = -\frac{1}{N_k} \sum_{n=1}^{N_k} \log \sigma\left(\tilde{r}_{k,n}\, \boldsymbol{w}_k^\top \boldsymbol{\phi}_k(\boldsymbol{x}_n)\right).$$

Minimizing ℓ_k encourages the model to assign higher weights to rules that support correct predictions and lower weights to those contributing to errors. To build and refine the rule set Γ_k, SEER employs a column generation (Dash et al., 2018) approach that alternates between:

- **Weight Update (Master Problem):** Given the current rule set Γ_k^m at iteration m, optimize rule weights $\boldsymbol{w}_k^m$ by minimizing $\ell_k(\boldsymbol{w}_k)$ as above.
- **Rule Expansion (Subproblem):** Identify the most informative new rule γ_k^{m+1} to add, by selecting the candidate rule γ not yet in Γ_k^m that maximizes the magnitude of the gradient of the loss w.r.t. its weight:

$$\gamma_k^{m+1} = \arg \max_{\gamma \in \Gamma \setminus \Gamma_k^m} \left| \frac{\partial \ell_k}{\partial w_\gamma} \right|,$$

where $\frac{\partial \ell_k}{\partial w_\gamma} = \frac{1}{N_k} \sum_{n=1}^{N_k} \frac{-\tilde{r}_{k,n}\, \phi_\gamma(\boldsymbol{x}_n)}{1+\exp\left(\tilde{r}_{k,n} \boldsymbol{w}_k^{m\top} \boldsymbol{\phi}_k(\boldsymbol{x}_n)\right)}$.

The process stops when the largest gradient magnitude falls below a predefined threshold λ, indicating no additional rules significantly improve the model.

Constructing $p^{\text{AI}}(k \mid \boldsymbol{x})$ The multi-class probability $p^{\text{AI}}(k \mid \boldsymbol{x})$ is computed by combining the binary predictors with a softmax function:

$$p^{\text{AI}}(k \mid \boldsymbol{x}, \boldsymbol{w}, \Gamma) = \frac{\exp\left(\eta\, \boldsymbol{w}_k^\top \boldsymbol{\phi}_k(\boldsymbol{x})\right)}{\sum_{i=1}^{K} \exp\left(\eta\, \boldsymbol{w}_i^\top \boldsymbol{\phi}_i(\boldsymbol{x})\right)},$$

where $\eta > 0$ is a temperature parameter controlling confidence calibration. To ensure the rule accuracy and reliability, in practice, we can adopt the following strategies: 1) Rules are initially sourced from *expert knowledge* or updated from data via the CG algorithm. 2) Before integration, all rules undergo human expert validation for correctness and clinical relevance. Together, this framework balances interpretability, adaptability, and accuracy, enabling SEER to evolve its rule-based predictor reliably over time.

4.2. Bayesian Estimation of Confusion Matrix

SEER adopts a Bayesian framework to estimate an expert's confusion matrix $\boldsymbol{C}^{(h)}$ using only partial reward signals $r_s \in \{+1, -1\}$. Let i denote the (unknown) true label, and y_s the expert's prediction for the s-th instance. To handle

the uncertainty in label supervision, we model the reward likelihood as $P(r_s \mid i, y_s)$ and place a Dirichlet prior over each row of the confusion matrix $\boldsymbol{C}_{i,:}^{(h)}$.

Under this model, the posterior distribution over the confusion matrix becomes:

$$\boldsymbol{C}_{i,:}^{(h)} \mid \text{Data} \sim \text{Dirichlet}(\alpha_{i1} + \mathbb{E}[n_{i1}], \ldots, \alpha_{iK} + \mathbb{E}[n_{iK}]),$$

where α_{ij} are the Dirichlet prior parameters, and $\mathbb{E}[n_{ij}] = \sum_s P(i \mid r_s, y_s) \cdot \mathbb{I}(y_s = j)$ denotes the expected number of times the expert predicted label j when the true label is inferred to be i.

The maximum a posteriori (MAP) estimate of the confusion matrix is then:

$$\hat{\boldsymbol{C}}_{ij}^{(h)} = \frac{\alpha_{ij} + \mathbb{E}[n_{ij}]}{\sum_{k=1}^{K} \left(\alpha_{ik} + \mathbb{E}[n_{ik}]\right)}.$$

This Bayesian formulation seamlessly integrates partial feedback with prior beliefs, providing a robust and data-efficient estimate of expert reliability. A full derivation of this estimation procedure is provided in Appendix N.

5. Experiments

We evaluate SEER on three settings: *synthetic*, *semi-synthetic*, and *real-world* datasets. The semi-synthetic setup uses real rare disease cases with simulated experts; the real-world setting involves image classification with true human labels. SEER demonstrates strong performance in terms of diagnostic accuracy, robustness to noise, identification of cognitive regions, and online rule learning. For synthetic and semi-synthetic data, we consider the following baselines and evaluation metrics:

Baselines and Comparison We compare SEER's accuracy against ensemble methods, including sparsely gated Mixture of Experts (MoE) (Shazeer et al., 2017), GLAD (Whitehill et al., 2009), weighted/average voting, and probabilistic fusion (P+L) (Kerrigan et al., 2021). Appendix C provides detailed descriptions of these methods.

Diagnostic Evaluation Metrics We evaluate our framework using three metrics: 1) **Accuracy:** Fraction of correct diagnoses. We report both overall and per-class accuracy to assess performance across diagnostic categories. 2) **Reward:** Simulated patient feedback. A binary reward (+1 or -1) is sampled with probability $\sigma(v) = \frac{1}{1+e^{-v}}$, where v is the summed weight of satisfied ground-truth rules supporting the current diagnosis. 3) **Regret:** The difference between the cumulative reward of our model and that of an oracle with full rule knowledge.

In simulations, all metrics are reported. In semi-synthetic and real-world studies, we focus on accuracy due to the lack of an oracle. Additionally, our SEER framework adapts over time by updating its model based on previously received patient feedback. Detailed experimental results from synthetic and real-world studies are presented in Appendices D and K, respectively.

5.1. Semi-Synthetic Experiments with Gitelman Syndrome Patient Data

To improve realism, we use a private dataset of real *Gitelman syndrome* patient records (71 positive, 95 negative) and simulate AI and human experts based on established clinical guidelines (Blanchard et al., 2017; of Chinese Research Hospital Association et al., 2022). Each patient record includes five key diagnostic features: *Serum Potassium*, *Urine Potassium*, *pH*, *Bicarbonate*, and *High Blood Pressure*. The task is framed as binary classification using leave-one-out cross-validation. Six simulated doctors and one AI agent make rule-based predictions, following the same protocol as in the synthetic experiments (see Appendix J for details).

Results Figure 2 shows the performance of various methods on real-world patient data. Among them, **SEER** achieves the highest accuracy of 0.885 ± 0.010, demonstrating its robustness in handling the inherent noise and complexity of real-world medical datasets. This is attributed to SEER's mechanism of mitigating AI uncertainty through strategic expert selection. Conversely, **P+L** exhibits poor performance with an accuracy of 0.270 ± 0.101, indicating its inability to effectively manage noisy labels from individual doctors due to the lack of sequential expert engagement.

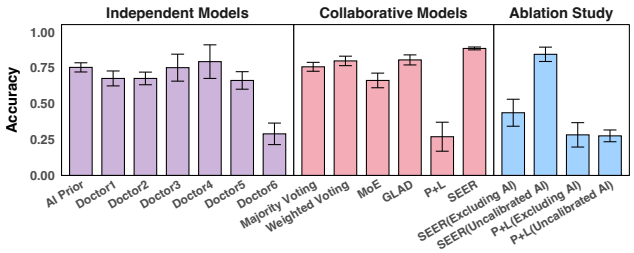

Figure 2. Bar plot comparing the performance of different models on the Gitelman syndrome dataset. Each bar represents the mean accuracy, with error bars indicating the standard deviation.

6. Conclusion

We introduced the Sequential Expert Engagement for Rare diseases (SEER), which dynamically integrates logic rule-based AI priors with expert insights through information gain, bridging human knowledge gaps in complex and rare scenarios. SEER enhances adaptability, interpretability, and robustness, consistently outperforming traditional methods across diverse datasets. Our theoretical analysis establishes conditions for optimal human-AI collaboration, while experiments on synthetic and real-world data validate SEER's ability to improve accuracy, resilience to noise, and expert selection efficiency.

Impact Statement

Our work introduces a novel human-AI collaborative decision-making framework that integrates logic rule-based AI with adaptive expert selection to optimize decision quality in high-stakes domains, especially the diagnosis of rare diseases. By leveraging the Sequential Expert Engagement for Rare diseases (SEER), our approach dynamically combines AI-generated insights with human expertise, ensuring interpretability and robustness. Empirical results demonstrate significant performance gains over the traditional ensemble and human-only methods, while theoretical analysis establishes conditions for optimal human-AI synergy. This research advances AI-augmented decision-making by improving efficiency, adaptability, and expert integration, enabling more effective and scalable applications in real-world settings.

References

Bagui, S. C. Combining pattern classifiers: Methods and algorithms. *Technometrics*, 47:517 – 518, 2005.

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., and Weld, D. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–16, 2021.

Baynam, G., Hartman, A. L., Letinturier, M. C. V., Bolz-Johnson, M., Carrion, P., Chen Grady, A., Dong, X., Dooms, M., et al. Global health for rare diseases through primary care. *Lancet Global Health*, 12(7):e1192–e1199, 2024. doi: 10.1016/S2214-109X(24)00134-7.

Blanchard, A., Bockenhauer, D., Bolignano, D., Calo, L. A., Cosyns, E., Devuyst, O., Ellison, D. H., Frankl, F. E. K., Knoers, N. V., Konrad, M., et al. Gitelman syndrome: consensus and guidance from a kidney disease: improving global outcomes (kdigo) controversies conference. *Kidney international*, 91(1):24–33, 2017.

Buçinca, Z., Swaroop, S., Paluch, A. E., Murphy, S. A., and Gajos, K. Z. Towards optimizing human-centric objectives in ai-assisted decision-making with offline reinforcement learning. *arXiv preprint arXiv:2403.05911*, 2024.

Chan, T.-H., Jia, K., Gao, S., Lu, J., Zeng, Z., and Ma, Y. Pcanet: A simple deep learning baseline for image classification? *IEEE transactions on image processing*, 24(12):5017–5032, 2015.

Chen, Y., Hassani, S. H., Karbasi, A., and Krause, A. Sequential information maximization: When is greedy near-optimal? In *Conference on Learning Theory*, pp. 338–363. PMLR, 2015.

Dash, S., Gunluk, O., and Wei, D. Boolean decision rules via column generation. *Advances in neural information processing systems*, 31, 2018.

Davis-Stober, C. P., Budescu, D. V., Broomell, S. B., and Dana, J. The composition of optimally wise crowds. *Decision Analysis*, 12(3):130–143, 2015.

De, A., Koley, P., Ganguly, N., and Gomez-Rodriguez, M. Regression under human assistance. *ArXiv*, abs/1909.02963, 2019.

De Toni, G., Okati, N., Thejaswi, S., Straitouri, E., and Gomez-Rodriguez, M. Towards human-ai complementarity with predictions sets. *arXiv preprint arXiv:2405.17544*, 2024a.

De Toni, G., Okati, N., Thejaswi, S., Straitouri, E., and Gomez-Rodriguez, M. Towards human-ai complementarity with predictions sets. *arXiv preprint arXiv:2405.17544*, 2024b.

Dietterich, T. G. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pp. 1–15. Springer, 2000.

Domaradzki, J. and Walkowiak, D. Medical students' knowledge and opinions about rare diseases: A case study from poland. *Intractable & rare diseases research*, 8 4:252–259, 2019.

Evans, W. Dare to think rare: diagnostic delay and rare diseases. *British Journal of General Practice*, 68 670: 224–225, 2018.

Finn, C. B., Tong, J. K., Alexander, H. E., Wirtalla, C., Wachtel, H., Guerra, C. E., Mehta, S. J., Wender, R., and Kelz, R. R. How referring providers choose specialists for their patients: a systematic review. *Journal of General Internal Medicine*, 37:3444–3452, 2022. doi: 10.1007/s11606-022-07574-6.

Genest, C. and Zidek, J. V. Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, 1(1):114–135, 1986.

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

Hong, L. and Page, S. E. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, 101 (46):16385–16389, 2004.

Kerrigan, G., Smyth, P., and Steyvers, M. Combining human predictions with model probabilities via confusion matrices and calibration. *Advances in Neural Information Processing Systems*, 34:4421–4434, 2021.

Kittler, J., Hatef, M., Duin, R. P., and Matas, J. On combining classifiers. *IEEE transactions on pattern analysis and machine intelligence*, 20(3):226–239, 1998.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Kurvers, R. H. J. M., Nuzzolese, A. G., Russo, A., Barabucci, G., Herzog, S. M., and Trianni, V. Automating hybrid collective intelligence in open-ended medical diagnostics. *Proceedings of the National Academy of Sciences*, 120(34):e2221473120, 2023. doi: 10.1073/pnas.2221473120.

Lamb, B. W., Sevdalis, N., Taylor, C., Vincent, C. A., and Green, J. S. A. Multidisciplinary team working across different tumour types: analysis of a national survey. *Annals of oncology : official journal of the European Society for Medical Oncology*, 23 5:1293–300, 2012.

Lamberson, P. and Page, S. E. Optimal forecasting groups. *Management Science*, 58(4):805–810, 2012.

Lee, M. D. and Lee, M. N. The relationship between crowd majority and accuracy for binary decisions. *Judgment and Decision Making*, 12(4):328–343, 2017.

Lee, Y.-J., Hosanagar, K., and Tan, Y. Do i follow my friends or the crowd? information cascades in online movie ratings. *Management Science*, 61(9):2241–2258, 2015.

Li, A. C., Prabhudesai, M., Duggal, S., Brown, E., and Pathak, D. Your diffusion model is secretly a zero-shot classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2206–2217, 2023.

Liu, H., Lai, V., and Tan, C. Understanding the effect of out-of-distribution examples and interactive explanations on human-ai decision making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–45, 2021.

Madras, D., Pitassi, T., and Zemel, R. S. Predict responsibly: Improving fairness and accuracy by learning to defer. In *Neural Information Processing Systems*, 2017.

Mozannar, H., Lang, H., Wei, D., Sattigeri, P., Das, S., and Sontag, D. Who should predict? exact algorithms for learning to defer to humans. In *International conference on artificial intelligence and statistics*, pp. 10520–10545. PMLR, 2023.

of Chinese Research Hospital Association, R. D. S., Group, G. S. C. W., et al. Expert consensus for the diagnosis and treatment of gitelman syndrome in china (2021). *Journal of Rare Diseases*, 1(1):56–67, 2022.

Pradier, M. F., Zazo, J., Parbhoo, S., Perlis, R. H., Zazzi, M., and Doshi-Velez, F. Preferential mixture-of-experts: Interpretable models that rely on human expertise as much as possible. *AMIA Summits on Translational Science Proceedings*, 2021:525, 2021.

Qian, M., xia Zhan, Y., Ji, L., and Cheng, Y. Strategy on precision medicine multidisciplinary team in clinical practice. *Clinical and Translational Discovery*, 2023.

Raman, N. and Yee, M. Improving learning-to-defer algorithms through fine-tuning. *arXiv preprint arXiv:2112.10768*, 2021.

Sagi, O. and Rokach, L. Ensemble learning: A survey. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 8(4):e1249, 2018.

Shaw, J., Rudzicz, F., Jamieson, T., and Goldfarb, A. Artificial intelligence and the implementation challenge. *Journal of medical Internet research*, 21(7):e13659, 2019.

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., and Dean, J. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Steyvers, M., Tejeda, H., Kerrigan, G., and Smyth, P. Bayesian modeling of human–ai complementarity. *Proceedings of the National Academy of Sciences*, 119(11):e2111547119, 2022.

The Lancet. Hope for rare diseases. *The Lancet*, 404(10464):1701, 2024. doi: 10.1016/S0140-6736(24)02414-0.

Venus, K., Kwan, J. L., and Frost, D. W. Rising to the challenge of rare diagnoses. *Journal of General Internal Medicine*, 40:918–921, 2025. doi: 10.1007/s11606-024-09086-x.

Verma, R., Barrejón, D., and Nalisnick, E. Learning to defer to multiple experts: Consistent surrogate losses, confidence calibration, and conformal ensembles. In *International Conference on Artificial Intelligence and Statistics*, pp. 11415–11434. PMLR, 2023.

Wang, W., Gao, G., and Agarwal, R. Friend or foe? teaming between artificial intelligence and workers with variation in experience. *Management Science*, 70(9):5753–5775, 2024.

Whitehill, J., Wu, T.-f., Bergsma, J., Movellan, J., and Ruvolo, P. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise.

Advances in neural information processing systems, 22, 2009.

Wilder, B., Horvitz, E., and Kamar, E. Learning to complement humans. In *International Joint Conference on Artificial Intelligence*, 2020.

Winters, D. A., Soukup, T., Sevdalis, N., Green, J. S., and Lamb, B. W. The cancer multidisciplinary team meeting: in need of change? history, challenges and future perspectives. *BJU International*, 128, 2021. URL https://api.semanticscholar.org/CorpusID:235168457.

Xu, L., Krzyzak, A., and Suen, C. Y. Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE transactions on systems, man, and cybernetics*, 22(3):418–435, 1992.

Yin, J., Ngiam, K. Y., Tan, S. S.-L., and Teo, H. H. Designing ai-based work processes: How the timing of ai advice affects diagnostic decision making. *Management Science*, 2025.

Zang, Z., Li, S., Wu, D., Wang, G., Wang, K., Shang, L., Sun, B., Li, H., and Li, S. Z. Dlme: Deep local-flatness manifold embedding. In *European Conference on Computer Vision*, pp. 576–592. Springer, 2022.

A. Related Work

Ensembles and Opinion Pools Prior research has convincingly demonstrated the performance advantages of leveraging multiple predictors over a single predictor. This principle is evident in both model combinations (Kittler et al., 1998; Bagui, 2005; Sagi & Rokach, 2018) and human opinion aggregations (Hong & Page, 2004; Lamberson & Page, 2012). Majority voting (Dietterich, 2000) and naive Bayes aggregation (Xu et al., 1992) are prevalent methods for aggregating non-probabilistic classifiers. However, majority voting may fall short in accuracy enhancement with a limited number of predictors, and naive Bayes aggregation, while effective at the class level, does not fully exploit instance-level uncertainties presented by probabilistic labelers. In the realm of human opinion ensembling, methods range from additive linear and log-linear opinion pools for subjective distributions (Genest & Zidek, 1986), to techniques for weighting linear combinations of continuous human predictions (Davis-Stober et al., 2015), and voting strategies for consolidating label predictions from multiple human predictors (Lee & Lee, 2017). Additionally, Whitehill et al. (2009) introduced the GLAD model, which jointly estimates annotator expertise and task difficulty, although its practical application is limited by identifiability challenges. Our work bridges this gap by developing a unified framework for the adaptive fusion of probabilistic machine outputs with categorical human decisions. SEER employs information-theoretic criteria to dynamically optimize the collaboration policy and estimate the expertise of human annotators, thereby enhancing overall predictive performance.

Human-AI Complementarity Human-AI Complementarity aims to enhance the accuracy of predictions made by human experts utilizing decision support systems beyond the capabilities of the experts alone or the AI classifiers independently (De et al., 2019; De Toni et al., 2024a). Despite this goal, empirical studies have frequently found that human-AI teams do not surpass the highest performance of either the human or the AI alone, even with AI explanations (Bansal et al., 2021; Liu et al., 2021). Several works model this challenge as a mixture of experts involving both humans and AI. This approach was initially introduced by (Madras et al., 2017) and later adapted by (Wilder et al., 2020) and (Pradier et al., 2021) with the introduction of a mixture of expert surrogates. However, these methods have often failed empirically due to difficulties in optimizing the loss function. Subsequent approaches have sought to improve these models, notably by enhancing calibration (Raman & Yee, 2021). Furthermore, (Buçinca et al., 2024) introduced offline reinforcement learning to develop decision support policies that optimize human-centric objectives, achieving improvements in joint human-AI accuracy. Nevertheless, these methods were not designed for contexts requiring collaboration between multiple humans and AI, thus overlooking the diversity in human groups. To address this gap, (Verma et al., 2023) introduced a model with ensemble prediction combining AI and human predictions, but optimization of collaboration costs is lacking. (Mozannar et al., 2023) formulated a novel surrogate loss function capable of deferring to one of the multiple users without combining AI and human predictions. Furthermore, (Steyvers et al., 2022) employed a Bayesian framework that combines human and machine predictions for better accuracy, yet it neglects interpretability and interaction design. Although (De Toni et al., 2024b) proposed a greedy method that achieves near-optimal prediction sets for a single human-AI team, they do not address the classifier's role or the interpretability of decisions, making their approach less applicable to real-world scenarios. In contrast, our work supports collaboration among multiple human experts and an AI agent, explicitly models cost-aware expert invitations and ensures interpretability in both the AI's guidance and the final collective decision.

B. Proof of Theorems

B.1. Proof of Theorem. 3.1

Proof. **Case 1: High-Confidence Regime**

1. *Entropy Before Consultation:*

$$H(p_{\text{comb}}^{(t)}) \leq -\underbrace{(1-\epsilon)\log(1-\epsilon)}_{\approx 0} - \underbrace{\epsilon \log \epsilon}_{\approx 0} \approx 0.$$

This holds because the entropy $H(p) = -\sum p \log p$, and for near-certainty distributions, the dominant term $(1-\epsilon)\log(1-\epsilon) \approx 0$ as $\epsilon \ll 1$. Additionally, the minor term $\epsilon \log \epsilon \approx 0$ as $\epsilon \to 0$

2. *Expected Posterior Entropy:*

$$\mathbb{E}[H(p_{\text{comb}}^{(t+1)})] \leq \underbrace{(1-\delta)\cdot 0}_{\text{Correct report}} + \underbrace{\delta \cdot H_{\max}}_{\text{Incorrect report}}$$
$$\approx \delta \log K \ll \log K.$$

The above inequalities hold because if $Y_h = y^*$, $p_{\text{comb}}^{(t+1)}(y^*) \propto (1-\epsilon)\cdot(1-\delta) \approx 1$. The posterior becomes even more concentrated on y^*, leading to entropy ≈ 0. In another case, suppose $Y_h = k \neq y^*$, then $p_{\text{comb}}^{(t+1)}(y^*) \propto (1-\epsilon)\cdot\delta \approx 0$. The posterior spread uncertainty across all K classes leads to maximum entropy $H_{\max} = \log K$. The expectation of all possible human responses simplifies to the correct report term and the incorrect report term.

3. *Information Gain:*

$$\Delta H_{h_{\text{reinforce}}} \approx 0 - \delta \log K.$$

Notice the negative sign, and non-experts with higher $\boldsymbol{C}_{y^*,k}^{(h)}$ ($k \neq y^*$) would produce larger δ, reducing information gain.

Case 2: Low-Confidence Regime

1. *Entropy Before Consultation:*

$$H(p_{\text{comb}}^{(t)}) \approx \log 2 = 1 \quad \text{(bits)}.$$

2. *Posterior Entropy Reduction:* For expert $h_{\text{disambiguate}}$:

$$H(p_{\text{comb}}^{(t+1)}|Y_h) \approx \begin{cases} 0 & \text{if } Y_h = y_1 \\ 0 & \text{if } Y_h = y_2. \end{cases}$$

3. *Information Gain:*

$$\Delta H_{h_{\text{disambiguate}}} = 1 - 0 = 1.$$

Non-experts would leave residual entropy $H_{\text{mid}} > 0$, yielding $\Delta H_h < 1$.

Generalization to $K > 2$: For multiclass problems, the algorithm:

1. Identifies top candidates y_1, y_2 through $p_{\text{comb}}^{(t)}$
2. Selects experts maximizing $\min(\boldsymbol{C}_{y_1,y_1}^{(h)}, \boldsymbol{C}_{y_2,y_2}^{(h)})$

Our algorithm maximizes mutual information $I(y^*; Y_h|p_{\text{comb}}^{(t)})$ by focusing on the dominant uncertain classes analogous to optimal Bayesian experimental design.

Case 3: Shallow but Informative Prior

We analyze the behavior of the expert selection criterion under a prior that weakly favors a particular class $y^* \in [K]$, which we interpret as the true label. The prior over y is defined as:

$$p^{(t)}(y) = \begin{cases} \frac{1}{K} + \eta & \text{if } y = y^*, \\ \frac{1}{K} - \frac{\eta}{K-1} & \text{otherwise,} \end{cases} \quad \text{where } \eta > 0 \text{ and } \eta \ll 1.$$

This prior is valid since it sums to 1:

$$\left(\frac{1}{K} + \eta\right) + (K-1)\left(\frac{1}{K} - \frac{\eta}{K-1}\right) = 1.$$

The goal is to select the expert $h \in \mathcal{H}$ that maximizes expected information gain:

$$\Delta H_h := H(p^{(t)}) - \mathbb{E}_{Y_h \sim p_h}[H(p^{(t+1)} \mid Y_h)],$$

where $p_h(k) = \Pr(Y_h = k) = \sum_y p^{(t)}(y)\, \boldsymbol{C}^{(h)}_{y,k}$ is the predicted distribution over the expert's response, and the posterior is updated by Bayes' rule:

$$p^{(t+1)}(y \mid Y_h = k) = \frac{p^{(t)}(y)\, \boldsymbol{C}^{(h)}_{y,k}}{\sum_{y'} p^{(t)}(y')\, \boldsymbol{C}^{(h)}_{y',k}}.$$

1. *Entropy of Prior.* Let $\epsilon_y := p^{(t)}(y) - \frac{1}{K}$. Then $\epsilon_{y^*} = \eta$, and $\epsilon_{y \neq y^*} = -\frac{\eta}{K-1}$. Using a second-order Taylor approximation of entropy around the uniform distribution:

$$H(p^{(t)}) \approx \log K - \frac{1}{2} \sum_y \frac{\epsilon_y^2}{1/K} = \log K - \frac{K}{2} \sum_y \epsilon_y^2.$$

   Compute the perturbation energy:

$$\sum_y \epsilon_y^2 = \eta^2 + (K-1)\left(\frac{\eta}{K-1}\right)^2 = \eta^2 \left(1 + \frac{1}{K-1}\right) = \eta^2 \cdot \frac{K}{K-1}.$$

   Therefore:

$$H(p^{(t)}) = \log K - \frac{K^2}{2(K-1)}\eta^2 + o(\eta^2).$$

2. *Expert Response Distribution.* Assume expert h satisfies $\boldsymbol{C}^{(h)}_{y^*,y^*} \geq 1 - \delta$ for small δ. Then:

$$p_h(y^*) = \sum_y p^{(t)}(y)\, \boldsymbol{C}^{(h)}_{y,y^*} \geq \left(\frac{1}{K} + \eta\right)(1-\delta) + \sum_{y \neq y^*} \left(\frac{1}{K} - \frac{\eta}{K-1}\right) \boldsymbol{C}^{(h)}_{y,y^*}.$$

   If $\boldsymbol{C}^{(h)}_{y,y^*} \ll 1$ for $y \neq y^*$, then $p_h(y^*)$ is significantly larger than $1/K$.

3. *Posterior Sharpness.* When $Y_h = y^*$, the posterior becomes:

$$p^{(t+1)}(y \mid Y_h = y^*) = \frac{p^{(t)}(y)\, \boldsymbol{C}^{(h)}_{y,y^*}}{p_h(y^*)}.$$

   Since $p^{(t)}(y^*)$ is slightly larger than uniform and $\boldsymbol{C}^{(h)}_{y^*,y^*} \geq 1 - \delta$, the posterior concentrates on y^*. Thus, the entropy of this posterior is strictly smaller than the prior, with the drop scaling with η and δ.

4. *Comparative Information Gain.* Suppose another expert h' has $\boldsymbol{C}^{(h')}_{y^*,y^*} \approx \frac{1}{K}$ (i.e., near-uniform). Then:

   - $p_{h'}(y^*) \approx \frac{1}{K}$, so the posterior is almost unchanged.
   - Entropy reduction is minimal for h', i.e., $\Delta H_{h'} \ll \Delta H_h$.

Maximizing expected information gain will select the expert h^* such that:

$$h^* = \arg\max_h \Delta H_h \quad \Rightarrow \quad \boldsymbol{C}^{(h^*)}_{y^*,y^*} \geq 1 - \delta,$$

for some small $\delta > 0$, whenever the prior slightly favors y^* (i.e., $\eta > 0$). Thus, SEER naturally selects experts most confident and accurate on the true label y^*, even under weak belief.

□

B.2. Proof of Theorem. 3.2

Sketch of Proof. **Step 1: Informative AI Prior Anchors Posterior.**

The initial combined posterior $p^{(0)}_{\text{comb}} = p^{\text{AI}}$ satisfies

$$p^{(0)}_{\text{comb}}(Y^*) \geq \frac{1}{K} + \eta > \frac{1}{K},$$

which implies the entropy

$$H(p^{(0)}_{\text{comb}}) < \log K,$$

providing a non-uniform "anchor" towards the true label Y^*.

Step 2: Expected Entropy Reduction per Round.

Each expert h has confusion matrix $C^{(h)}$ with diagonal dominance on Y^*:

$$C^{(h)}_{Y^*,Y^*} > C^{(h)}_{k,Y^*} \quad \forall k \neq Y^*.$$

This implies the KL divergence between the expert's distribution over responses and uniform guessing is positive:

$$D_{\text{KL}}\big(C^{(h)}_{Y^*,\cdot} \| \mathcal{U}\big) > 0.$$

Selecting the expert h_t that maximizes expected entropy reduction ΔH_h ensures:

$$\gamma := \min_h \Delta H_h > 0,$$

so that

$$\mathbb{E}[H(p^{(t+1)}_{\text{comb}})] \leq H(p^{(t)}_{\text{comb}}) - \gamma,$$

guaranteeing at least a constant expected drop in entropy each round.

Step 3: Linear Improvement of Log-Odds.

Define log-odds:

$$\mathcal{L}_t = \log \frac{p^{(t)}_{\text{comb}}(Y^*)}{1 - p^{(t)}_{\text{comb}}(Y^*)}.$$

Because the experts are partially correct and chosen to maximize information gain, the expected increment in log-odds satisfies

$$\mathbb{E}[\mathcal{L}_{t+1} \mid \mathcal{L}_t] \geq \mathcal{L}_t + \Delta_{\min},$$

for some $\Delta_{\min} > 0$ that depends on the minimal KL divergence of expert confusion matrices from uniform guessing.

By induction:

$$\mathbb{E}[\mathcal{L}_t] \geq \mathcal{L}_0 + t\Delta_{\min},$$

which implies $p^{(t)}_{\text{comb}}(Y^*) \to 1$ exponentially fast in expectation.

Step 4: Combining Results and Complexity Bound.

Since entropy is a concave function decreasing to zero as the posterior mass on Y^* approaches 1, the linear growth of log-odds implies exponential entropy decay:

$$\mathbb{E}[H(p^{(T)}_{\text{comb}})] \leq H(p^{(0)}_{\text{comb}}) - T\gamma.$$

Moreover, to achieve posterior confidence $p^{(T)}_{\text{comb}}(Y^*) \geq \tau$, solve

$$\log \frac{\tau}{1-\tau} \leq \mathbb{E}[\mathcal{L}_T] \leq \mathcal{L}_0 + T\Delta_{\min} \implies T \geq \frac{\log \frac{\tau}{1-\tau} - \mathcal{L}_0}{\Delta_{\min}}.$$

This gives an explicit consultation complexity bound.

□

C. Comparisons with Ensemble and Majority Voting

Distinction from Ensemble Our algorithm's sequential human selection and querying scheme differs fundamentally from pure probability calibration, which adjusts probabilities to align with ground-truth frequencies (e.g., scaling AI confidence scores to match empirical accuracy). Specifically, our algorithm leverages the current belief state p_{comb} to dynamically select the next human expert h_t, guided by information gain and confusion matrix. Information gain selects h_t to minimize the entropy of p_{comb}, and confusion matrix prioritizes experts whose historical performance $\boldsymbol{C}^{(h_t)}$ indicates that they can resolve ambiguities in p_{comb}.

Comparison with Majority Voting Our algorithm can outperform majority voting under various conditions: when the AI's prior is strong, it reduces the need for human consultations needed to reach a target confidence level; when high-quality experts are rare, the algorithm strategically prioritizes the best experts, avoiding dilution of their input; when human errors are correlated, the algorithm mitigates biases by choosing experts with uncorrelated confusion matrices.

Table 1. Comparison of different models in the synthetic experiments (mean ± std over 10 runs). **Blue shading** indicates the best *independent* model; **Red shading** indicates the best *collaborative* model.

Methods	Accuracy ↑				Rewards ↑	Regret ↓
	Overall	Class-0	Class-1	Class-2		
Independent Models						
AI Prior	0.510 ± 0.028	0.465 ± 0.061	0.509 ± 0.054	0.558 ± 0.067	65.900 ± 3.833	17.000 ± 5.441
Doctor1	0.491 ± 0.062	0.544 ± 0.070	**0.558 ± 0.080**	0.370 ± 0.065	65.600 ± 5.024	17.300 ± 6.165
Doctor2	0.499 ± 0.038	0.538 ± 0.072	0.364 ± 0.065	**0.594 ± 0.097**	64.400 ± 3.412	18.500 ± 6.830
Doctor3	**0.559 ± 0.049**	0.576 ± 0.070	0.570 ± 0.065	0.530 ± 0.088	66.400 ± 2.691	16.500 ± 4.201
Doctor4	0.524 ± 0.029	0.571 ± 0.055	0.527 ± 0.074	0.473 ± 0.080	64.800 ± 3.789	18.100 ± 6.107
Doctor5	0.522 ± 0.053	**0.621 ± 0.073**	0.409 ± 0.101	0.533 ± 0.079	**68.900 ± 6.580**	**14.000 ± 10.050**
Doctor6	0.525 ± 0.037	0.497 ± 0.068	0.485 ± 0.030	0.594 ± 0.049	67.000 ± 4.837	15.900 ± 8.479
Doctor7	0.458 ± 0.035	0.388 ± 0.076	0.455 ± 0.078	0.533 ± 0.086	63.200 ± 4.686	19.700 ± 5.849
Doctor8	0.419 ± 0.048	0.347 ± 0.080	0.500 ± 0.071	0.412 ± 0.077	63.200 ± 4.069	19.700 ± 6.404
Doctor9	0.365 ± 0.050	0.367 ± 0.059	0.403 ± 0.085	0.324 ± 0.062	62.800 ± 4.729	20.100 ± 7.203
Collaborative Models						
Majority Voting	0.678 ± 0.038	0.671 ± 0.066	0.673 ± 0.085	0.691 ± 0.047	70.200 ± 3.709	12.700 ± 7.376
Weighted Voting	0.677 ± 0.041	0.682 ± 0.061	0.688 ± 0.072	0.661 ± 0.056	72.000 ± 3.688	10.900 ± 4.888
GLAD	0.262 ± 0.037	0.271 ± 0.074	0.273 ± 0.092	0.242 ± 0.094	58.200 ± 5.793	24.700 ± 8.210
MoE	0.669 ± 0.054	0.685 ± 0.086	0.630 ± 0.091	0.691 ± 0.048	72.600 ± 2.107	10.300 ± 4.981
P+L	0.707 ± 0.051	0.665 ± 0.103	0.712 ± 0.111	0.745 ± 0.049	75.200 ± 2.857	7.700 ± 5.622
SEER	**0.784 ± 0.035**	**0.800 ± 0.090**	**0.791 ± 0.091**	**0.758 ± 0.042**	**76.800 ± 2.960**	**6.100 ± 1.700**
Ablation Study						
SEER(Excluding AI)	0.559 ± 0.054	0.574 ± 0.059	0.536 ± 0.088	0.567 ± 0.124	69.200 ± 4.285	13.700 ± 6.198
SEER(Uncalibrated AI)	0.767 ± 0.043	0.785 ± 0.086	0.758 ± 0.107	0.730 ± 0.087	73.700 ± 2.722	9.200 ± 5.192
P+L(Excluding AI)	0.707 ± 0.051	0.665 ± 0.103	0.712 ± 0.111	0.745 ± 0.049	70.700 ± 5.178	12.200 ± 5.706
P+L(Uncalibrated AI)	0.707 ± 0.051	0.665 ± 0.103	0.712 ± 0.111	0.745 ± 0.049	74.100 ± 4.277	8.800 ± 6.660

D. Results of Synthetic Experiments

D.1. Synthetic Experiment: Simulating Rare Disease Diagnosis, Cognitive Gaps, and Rule-Based Agents

We conducted synthetic experiments to simulate rare disease diagnosis scenarios, generating datasets with predefined rules (see Table 2). Nine rule-based simulated doctors and an AI agent with a general cognitive domain were created. The AI agent, knowledgeable about common and rare diseases, offers a broad perspective but lacks the precision of specialists like Doctors 1-2 (experts in common and rare diseases), Doctors 4-5 (rare disease specialists), or Doctors 6-7 (mixed experts). Doctors 8-9 act as noisy experts, while Doctor 3, a comprehensive decision-maker, outperforms the AI agent. We produced a training set of 20,000 samples with $\{\mathbf{x}_t, y_{l=1}^L, k_t, r_t\}$ and an evaluation set of 100 samples with only patient features $\mathbf{x}_t$. Each method was trained on the training set and tested on the evaluation set. Unlike SEER, *other methods directly incorporate all decision-makers*: weighted and majority voting rely on voting outcomes for final decisions, while the remaining methods select the label with the highest probability. Further simulation details are provided in Appendix E.

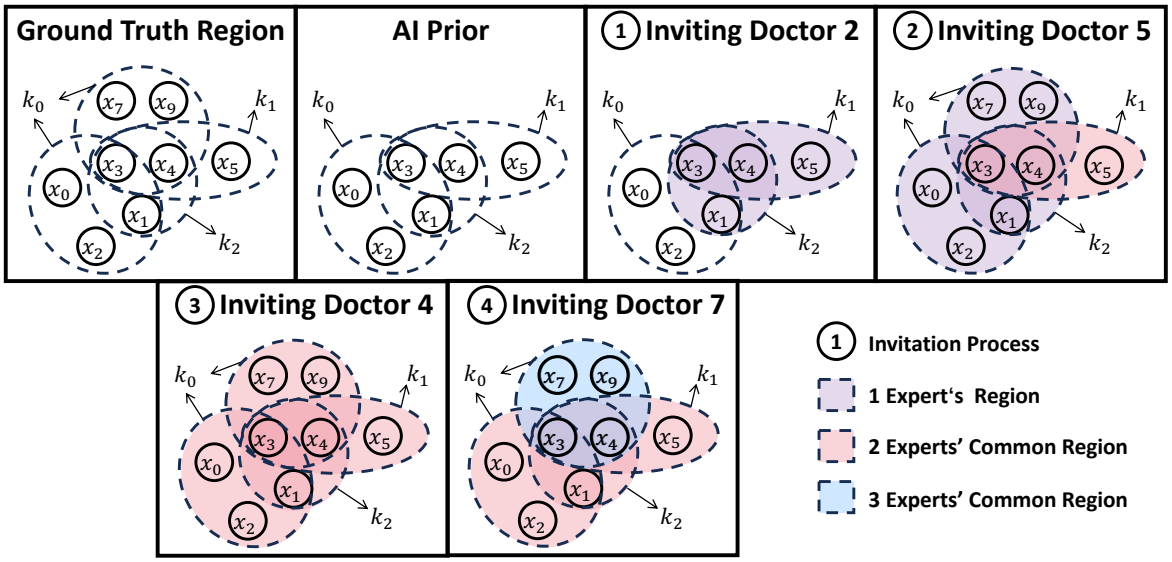

Figure 3. Dynamics of the cognitive region of human-AI team following sequential expert invitation.

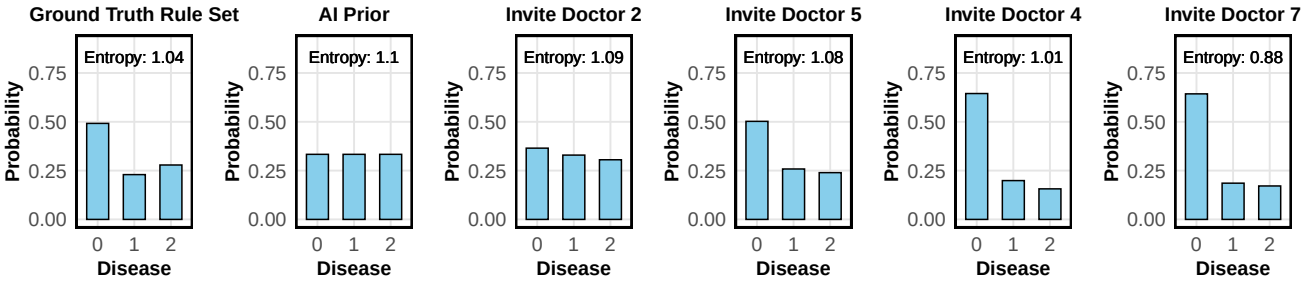

Figure 4. Comparison of probability distributions and entropy after each invitation of experts.

Results The experiments' results are shown in Table 1, SEER consistently outperforms all baselines across three diagnostic accuracy metrics. Our *ablation study* tests the role of AI priors in initializing SEER's belief. "Excluding AI" refers to removing the AI prior (i.e., randomly pick human experts), and "Uncalibrated AI" uses raw AI outputs without post-hoc calibration. We show that calibrating the AI's prior—e.g., using validation data (Guo et al., 2017) and aggregated expert feedback—can significantly improve decision quality.

To illustrate human-AI complementarity, we construct a challenging patient case that satisfies all rules for a rare disease (k_0) and one rule each for two common diseases (k_1, k_2) (details in Appendix F). Figure 3 shows how SEER sequentially recruits the most informative doctors to recover the correct cognitive region, reducing uncertainty (Figure 4). This highlights how

human experts can reinforce or disambiguate AI's general prior through complementary expertise. Appendix G confirms SEER's rule-learning behavior. Even when AI is weak (Appendix H), SEER maintains robustness by leveraging sequential information gain. In an idealized setting with no weak decision-makers (Appendix I), SEER still outperforms all baselines.

E. Setting of Synthetic Experiments

The framework integrates two principal entities: an AI agent and human agents. We assess their performance through a series of simulations. Below, we detail the architecture of our simulation framework.

Patient Simulator The sample generation process is based on the predefined set of ground truth rules, as shown in Table. 2. Initially, the generator selects a set of labels based on the rule weights, simulating population-level decision-making processes. This selection is formalized by the equation:

$$k \sim \text{Mult}\left(\text{softmax}\left(\text{sigmoid}^{-1}\left(\boldsymbol{w}_1^\top \boldsymbol{\phi}_1(\boldsymbol{x}), \ldots, \boldsymbol{w}_k^\top \boldsymbol{\phi}_k(\boldsymbol{x})\right)\right)\right),$$

where $\boldsymbol{w}_1^\top \boldsymbol{\phi}_1(\boldsymbol{x})$ represents the feature functions associated with the rule set, and the labels are selected if the corresponding features satisfy the rule conditions. Following label selection, we generate a random binary sequence for the patient. For a valid patient sample, at least one of the rules corresponding to the selected labels must be satisfied, while none of the rules associated with other labels should hold. In our simulation, label k_0 represents a rare disease, which typically requires more complex diagnostic processes, such as genetic testing, and thus is governed by longer rules. In contrast, labels k_1 and k_2 correspond to common diseases with simpler diagnostic criteria.

We divided the entire dataset into two disjoint subsets: (i) a training dataset $\mathcal{D}_t$ with 20000 samples and (ii) an evaluation dataset $\mathcal{D}_e$ with 100 samples. The training dataset $\mathcal{D}_t = \{\boldsymbol{x}_t, \{y_l\}_{l=1}^L, y_t, r_t, y^*\}$ includes comprehensive data such as consultation history and patient feedback. The evaluation dataset $\mathcal{D}_e$ is exclusively reserved for evaluation purposes.

Table 2. The ground truth rule set.

Label	Rules	Weight
k_0	1: $x_0 \wedge x_1 \wedge \neg x_2 \wedge x_3$	1.5
	2: $x_3 \wedge x_4 \wedge x_7 \wedge \neg x_9$	1.5
k_1	3: $x_3 \wedge x_4 \wedge x_5$	1.4
	4: $x_6 \wedge x_7 \wedge x_9$	1.6
k_2	5: $x_1 \wedge x_3 \wedge x_4$	1.7
	6: $x_4 \wedge x_7 \wedge x_9$	1.3

Human Doctor Simulator In real-world healthcare environments, doctors from various specialties exhibit distinct domains of expertise and may even hold inaccurate beliefs. To simulate this, we model each human doctor as a rule-based probabilistic decision-maker equipped with a unique set of rules that reflect their specific expertise and biases. These rule sets can demonstrate preferences for specific treatments or deviate significantly from the established ground-truth rule set. Unlike real-world scenarios where doctors typically provide deterministic treatment choices, our simulated doctors select the treatment corresponding to the probability distribution from the softmax function.

Our simulation framework includes a diverse pool of nine doctors. For each patient scenario, the AI agent first offers a treatment suggestion. If the entropy of the collective treatment distribution does not meet a predefined threshold (set at 0.3 in all our experiments), indicating a lack of consensus or insufficient confidence among the initial doctors, the SEER algorithm is employed. The SEER algorithm then invites additional doctors from the pool to contribute their recommendations, aiming to refine the decision-making process and enhance the reliability of the treatment choice. Details of all ten rule-based decision-makers can be found in Table 3.

AI Agent Learning Our AI agent, initially configured with a subset of ground truth rules, operates as a rule-based probabilistic decision-maker and serves as a general expert. This configuration mimics real-world scenarios where prior knowledge informs decision-making frameworks. Throughout the simulation, the AI agent dynamically updates its rule set and associated weights based on incoming data to refine its decision-making process.

Table 3. The rules assigned to each rule-based decision-maker.

Model	Rule Set	Weight	Model	Rule Set	Weight
AI Agent	$a_0 \leftarrow x_0 \wedge x_1 \wedge \neg x_2 \wedge x_3$	1.5	Human Doctor 1	$a_1 \leftarrow x_3 \wedge x_4 \wedge x_5$	1.4
	$a_1 \leftarrow x_3 \wedge x_4 \wedge x_5$	1.5		$a_1 \leftarrow x_6 \wedge x_7 \wedge x_9$	1.6
	$a_2 \leftarrow x_1 \wedge x_3 \wedge x_4$	1.5		$a_2 \leftarrow x_1 \wedge x_3 \wedge x_4$	1.7
				$a_0 \leftarrow x_3 \wedge x_4$	1.7
Human Doctor 2	$a_2 \leftarrow x_1 \wedge x_3 \wedge x_4$	1.7	Human Doctor 3	$a_1 \leftarrow x_3 \wedge x_4 \wedge x_5$	1.4
	$a_2 \leftarrow x_4 \wedge x_7 \wedge x_9$	1.3		$a_1 \leftarrow x_6 \wedge x_7 \wedge x_9$	1.6
	$a_1 \leftarrow x_3 \wedge x_4 \wedge x_5$	1.4		$a_2 \leftarrow x_1 \wedge x_3 \wedge x_4$	1.7
	$a_0 \leftarrow x_3 \wedge x_4$	1.4		$a_2 \leftarrow x_4 \wedge x_7 \wedge x_9$	1.3
				$a_0 \leftarrow x_3 \wedge x_4$	1.7
Human Doctor 4	$a_0 \leftarrow x_0 \wedge x_1 \wedge \neg x_2 \wedge x_3$	1.5	Human Doctor 5	$a_0 \leftarrow x_0 \wedge x_1 \wedge \neg x_2 \wedge x_3$	1.5
	$a_0 \leftarrow x_3 \wedge x_4 \wedge x_7 \wedge \neg x_9$	1.5		$a_0 \leftarrow x_3 \wedge x_4 \wedge x_7 \wedge \neg x_9$	1.5
	$a_1 \leftarrow x_3 \wedge x_4$	1.4		$a_2 \leftarrow x_3 \wedge x_4$	1.5
	$a_2 \leftarrow x_1 \wedge x_3 \wedge x_4$	1.7		$a_1 \leftarrow x_3 \wedge x_4 \wedge x_5$	1.7
Human Doctor 6	$a_0 \leftarrow x_0 \wedge x_1 \wedge \neg x_2 \wedge x_3$	1.5	Human Doctor 7	$a_0 \leftarrow x_3 \wedge x_4 \wedge x_7 \wedge \neg x_9$	1.5
	$a_2 \leftarrow x_1 \wedge x_3 \wedge x_4$	1.7		$a_1 \leftarrow x_6 \wedge x_7 \wedge x_9$	1.6
	$a_1 \leftarrow x_3 \wedge x_4 \wedge x_5$	1.5		$a_2 \leftarrow x_3 \wedge x_4$	1.3
Human Doctor 8	$a_0 \leftarrow x_0 \wedge \neg x_2$	1.5	Human Doctor 9	$a_0 \leftarrow x_3 \wedge x_4$	1.5
	$a_1 \leftarrow x_3 \wedge x_4$	1.5		$a_1 \leftarrow x_3 \wedge x_4$	1.5
	$a_2 \leftarrow x_1 \wedge x_3$	1.5		$a_2 \leftarrow x_4 \wedge x_7 \wedge x_9$	1.5

To facilitate the learning of these rules, our system relies on a reward feedback mechanism. Specifically, we simulate an oracle environment using the ground truth rule set. For each patient data instance, the human-AI collaboration team proposes a diagnosis, which is then evaluated by the oracle. The oracle assesses this diagnosis by computing the conditional probability of the treatment given the patient's data. Subsequently, a reward is sampled using a Bernoulli distribution based on this probability. This reward signal serves as crucial feedback, enabling the AI agent to optimize its rule set for improved decision accuracy over time.

The metrics employed to evaluate the performance of rule learning are the weights' mean absolute error (MAE) and rule accuracy. The calculation of weights' MAE follows a stringent method: for rules accurately identified in the ground-truth rule set, we directly compute the absolute error. For those not accurately identified, we assign the absolute error to be equal to the true weight. Regarding rule accuracy, if the rules in the ground-truth rule set are not accurately identified, we account for this in our evaluation.

F. Experimental Results: Visualization of decision making process

To demonstrate the decision-making process of our framework, we constructed a sample with the feature vector $[x_0 = 1, x_1 = 1, x_2 = 0, x_3 = 1, x_4 = 1, x_5 = 1, x_6 = 0, x_7 = 1, x_8 = 0, x_9 = 0]$, which lies outside the expertise of any single model within the framework. This sample satisfies the conditions for rules 1, 2, 3, and 5, which are associated with all three possible labels, thereby illustrating the necessity for a collaborative decision-making approach.

As illustrated in Figure 3, inviting more informative human experts significantly improves the recovery of the cognitive region of the human-AI team. This process demonstrates the collaborative strength of our framework in leveraging diverse expert inputs to enhance decision-making. Additionally, the entropy of the final distribution after calibration, depicted in Figure 4, is minimized, enabling a more accurate recovery of the ground truth conditional distribution.

G. Experimental Results: Rule Learning

We validate the rule-learning capability of our AI agent by simulating real-world scenarios with a cohort of 20,000 pre-generated patients. Initially, the AI agent provides its assessments and then invites experts based on the criterion of maximizing information gain. An oracle subsequently delivers feedback in the form of rewards, enabling us to update the AI agent's parameters every 5,000 patients. To assess the decision-making ability of our AI agent, we utilize accuracy and reward metrics, as illustrated in Figure 5. The mean absolute error (MAE) and standard deviation of the learned rule weights across ten replicates are reported in Table 4, alongside the accuracy of the learned rules in these replicates. The results demonstrate that our rule-learning method is highly effective in accurately identifying the ground-truth cognitive region with sufficient data. This validation shows that our AI agent can adapt and optimize its rule set as new data is incorporated,

further supporting its application in complex, real-world decision-making environments.

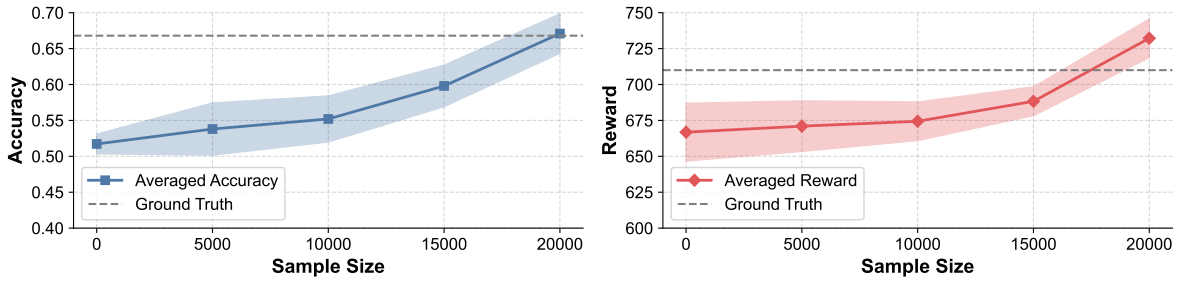

Figure 5. Learning curves for AI-agent on synthetic experiments. The x-axis represents the number of samples in the training dataset, while the y-axis shows the accuracy and the number of positive rewards in a total of 1000 evaluation samples. The shaded region represents the standard deviation.

We report the rule learning accuracy of rule content and the MAE of rule weight in Table. 4. Besides, we also report the agent's development by detailing the evolution of the rule set across varying training sample sizes in one of the repeated experiments, as shown in Table. 5.

Table 4. The rule learning accuracy of the AI agent across different sample sizes is illustrated here. Details on how to calculate rule accuracy and weight mean absolute error (MAE) can be found in Appendix. E.

Sample Size	Rule Accuracy	Weight MAE
5000	0.55 ± 0.08	0.774 ± 0.118
10000	0.55 ± 0.08	0.757 ± 0.117
15000	0.80 ± 0.07	0.382 ± 0.113
20000	0.98 ± 0.05	0.134 ± 0.065

H. Experimental Results: Specific Failure Modes

To address concerns about the limited scope of our experiments and to evaluate the robustness of our approach under less ideal conditions, we conducted an additional simulation designed to simulate a potential failure mode. In this scenario, the AI agent has access to only two out of the six rules from the ground truth rule set, with each rule assigned a random weight. This setup emulates a situation where the AI agent's knowledge is incomplete, reflecting real-world cases where the agent may operate with partial or noisy information.

The results for this simulation are summarized in Table 6. Under this challenging scenario, SEER's performance does not maintain the superior levels observed in previous experiments. However, it still outperforms alternative methods in accuracy, demonstrating its robustness and adaptability even in adverse settings. While some degradation in accuracy, rewards, and regret is observed compared to the original experiments, SEER consistently shows competitive results, outperforming many baseline approaches.

I. Experimental Results: Without Noisy Experts

We conducted an additional experiment by excluding noisy experts from prior simulations to evaluate the performance of various methods when only reliable expertise is available. The remaining experts exhibit relatively high overall accuracy, providing consistent and dependable input. This setup assesses the robustness and efficacy of each method in a noise-free environment.

Table 8 presents the evaluation results. Our SEER framework consistently outperformed other approaches, demonstrating its strength in integrating detailed distributions from experts with diverse expertise levels, rather than merely aggregating labels. Although all methods exhibited improved accuracy and rewards, as well as reduced regret, SEER's superior performance underscores its effectiveness in leveraging high-quality expert contributions.

Table 5. Evolution of rule sets and weights possessed by the AI agent after processing every 5,000 samples. This table reports all learned rules during each update.

Samples	Rule Sets	Weight
5000	$k_0 \leftarrow x_0 \wedge x_1 \wedge \neg x_2 \wedge x_3$	1.315
	$k_0 \leftarrow x_3 \wedge x_4 \wedge x_7 \wedge \neg x_9$	0.866
	$k_1 \leftarrow x_3 \wedge x_4 \wedge x_5$	1.663
	$k_2 \leftarrow x_1 \wedge x_3 \wedge x_4$	1.322
10000	$k_0 \leftarrow x_0 \wedge x_1 \wedge \neg x_2 \wedge x_3$	1.246
	$k_0 \leftarrow x_3 \wedge x_4 \wedge x_7 \wedge \neg x_9$	1.340
	$k_1 \leftarrow x_3 \wedge x_4 \wedge x_5$	1.673
	$k_2 \leftarrow x_1 \wedge x_3 \wedge x_4$	1.010
15000	$k_0 \leftarrow x_0 \wedge x_1 \wedge \neg x_2 \wedge x_3$	1.256
	$k_0 \leftarrow x_3 \wedge x_4 \wedge x_7 \wedge \neg x_9$	1.336
	$k_1 \leftarrow x_3 \wedge x_4 \wedge x_5$	1.542
	$k_1 \leftarrow x_6 \wedge x_7 \wedge x_9$	1.339
	$k_2 \leftarrow x_1 \wedge x_3 \wedge x_4$	1.188
20000	$k_0 \leftarrow x_0 \wedge x_1 \wedge \neg x_2 \wedge x_3$	1.439
	$k_0 \leftarrow x_3 \wedge x_4 \wedge x_7 \wedge \neg x_9$	1.435
	$k_1 \leftarrow x_3 \wedge x_4 \wedge x_5$	1.445
	$k_1 \leftarrow x_6 \wedge x_7 \wedge x_9$	1.538
	$k_2 \leftarrow x_1 \wedge x_3 \wedge x_4$	1.661
	$k_2 \leftarrow x_4 \wedge x_7 \wedge x_9$	1.369

J. Details of Semi-Synthetic Experiments

J.1. Participants and Decision-Making Framework

- **Doctors:** Six human doctors participated, utilizing the following approaches:
  - **Doctor 1–3:** Operated within unique cognitive regions, leveraging subsets or modified versions of the overall rule set.
  - **Doctor 4:** Employed the comprehensive overall rule set.
  - **Doctor 5:** Utilized an alternate subset of rules.
  - **Doctor 6:** Followed a random decision-making strategy.
- **AI Agent:** Utilized three rules from the overall rule set.

J.2. Rule Sets

The comprehensive rule set used for diagnosing Gitelman syndrome is presented in Table 9. Each rule is assigned a weight, reflecting its relative importance.

K. Experimental Results: Real Human Annotation Data with CIFAR-10H

Dataset CIFAR-10H[1] contains soft labels capturing human perceptual uncertainty, with multiple annotations per image. We intentionally introduce noise to human labels (the controlled label corruption protocol systematically reduces accuracy to 0.816 ± 0.027 (mean ± std) by uniformly misassigning partially correct labels to erroneous categories through random re-assignment) to test our framework's ability to improve classification through human-AI collaboration.

Baseline AI Models To fairly compare the performance of our proposed framework, we consider four different models for image classification tasks: *(i) DCNN* (Krizhevsky et al., 2012), *(ii) PCANet* (Chan et al., 2015), and *(iii) DC (Diffusion Classifier)* (Li et al., 2023), and *(iv) DLME* (Zang et al., 2022). Furthermore, we also compared the human annotation accuracy after introducing noise. Our framework is consistent across all AI models.

[1] https://github.com/jcpeterson/cifar-10h

Table 6. Performance evaluation with one AI agent has incomplete knowledge (as shown in Table 7). Results are presented as mean ± standard deviation across 10 repeated runs. **Blue shading** indicates the best *independent* model; **Red shading** indicates the best *collaborative* model.

Methods	Accuracy ↑				Rewards ↑	Regret ↓
	Overall	Class-0	Class-1	Class-2		
Independent Models						
AI Prior	0.491 ± 0.058	0.426 ± 0.071	0.512 ± 0.105	0.536 ± 0.072	0.666 ± 0.046	0.142 ± 0.065
Doctor1	0.481 ± 0.042	0.482 ± 0.056	0.558 ± 0.115	0.403 ± 0.073	0.672 ± 0.037	0.136 ± 0.056
Doctor2	0.454 ± 0.028	0.497 ± 0.059	0.355 ± 0.082	0.509 ± 0.065	0.642 ± 0.031	0.166 ± 0.066
Doctor3	0.534 ± 0.053	0.506 ± 0.091	**0.606 ± 0.098**	0.491 ± 0.102	0.668 ± 0.050	0.140 ± 0.070
Doctor4	0.525 ± 0.034	**0.594 ± 0.055**	0.521 ± 0.083	0.458 ± 0.092	0.652 ± 0.037	0.156 ± 0.061
Doctor5	0.522 ± 0.047	0.585 ± 0.075	0.430 ± 0.084	0.548 ± 0.076	**0.695 ± 0.034**	**0.113 ± 0.048**
Doctor6	**0.542 ± 0.047**	0.503 ± 0.095	0.533 ± 0.073	**0.591 ± 0.087**	0.664 ± 0.044	0.144 ± 0.069
Doctor7	0.479 ± 0.037	0.406 ± 0.075	0.488 ± 0.048	0.545 ± 0.061	0.658 ± 0.042	0.150 ± 0.068
Doctor8	0.396 ± 0.053	0.326 ± 0.076	0.473 ± 0.049	0.391 ± 0.106	0.602 ± 0.036	0.206 ± 0.065
Doctor9	0.344 ± 0.042	0.368 ± 0.087	0.370 ± 0.052	0.294 ± 0.054	0.628 ± 0.037	0.180 ± 0.052
Collaborative Models						
Majority Voting	0.710 ± 0.035	**0.753 ± 0.077**	0.682 ± 0.094	0.694 ± 0.058	0.722 ± 0.044	0.086 ± 0.085
Weighted Voting	0.696 ± 0.035	0.724 ± 0.078	0.673 ± 0.067	0.691 ± 0.078	**0.746 ± 0.055**	**0.062 ± 0.059**
GLAD	0.281 ± 0.043	0.288 ± 0.093	0.309 ± 0.047	0.245 ± 0.044	0.603 ± 0.037	0.205 ± 0.059
MoE	0.662 ± 0.046	0.709 ± 0.062	0.609 ± 0.093	0.667 ± 0.052	0.703 ± 0.047	0.105 ± 0.080
P+L	0.546 ± 0.056	0.303 ± 0.095	**0.897 ± 0.134**	0.445 ± 0.218	0.678 ± 0.041	0.130 ± 0.058
SEER	**0.733 ± 0.056**	0.565 ± 0.132	0.858 ± 0.041	**0.782 ± 0.067**	0.714 ± 0.035	0.094 ± 0.063
Ablation Study						
SEER(Excluding AI)	0.529 ± 0.037	0.353 ± 0.102	0.779 ± 0.115	0.461 ± 0.101	0.670 ± 0.050	0.138 ± 0.071
SEER(Uncalibrated AI)	0.703 ± 0.076	0.571 ± 0.144	0.818 ± 0.080	0.724 ± 0.125	0.696 ± 0.066	0.112 ± 0.100
P+L(Excluding AI)	0.546 ± 0.056	0.303 ± 0.095	0.897 ± 0.134	0.445 ± 0.218	0.684 ± 0.040	0.124 ± 0.063
P+L(Uncalibrated AI)	0.546 ± 0.056	0.303 ± 0.095	0.897 ± 0.134	0.445 ± 0.218	0.663 ± 0.044	0.145 ± 0.069

Table 7. The rule set for our weak AI agent.

Label	Rules	Weight
k_0	1: $x_0 \wedge x_1 \wedge \neg x_2$	1.5
k_1	2: $x_3 \wedge x_4 \wedge x_5$	1.5
k_2	3: $x_1 \wedge x_3 \wedge x_4$	1.5

Results Figure 6 reveals two distinct patterns: While DCNN/PCANet underperform noisy human annotations (72.3-81.6%), DC/DLME surpass them (85.5-89.2%). Our framework consistently outperforms standalone AI models through strategic human collaboration. Traditional voting excels with reliable human input, but SEER better handles noise, particularly with a robust AI agent(DC/DLME), achieving 89.1-92.5% accuracy. Calibrated AI agents yield 3.8-4.3% confidence improvements over uncalibrated versions. Encouragingly, our framework also reduces expert annotation demands, requiring only 5 human experts per figure compared to the 10-expert requirement of baseline methods, while maintaining superior accuracy. Overall, these results validate our framework's practical effectiveness with real-world data.

L. Limitation and Broader Impacts

A notable limitation of our proposed model lies in the rule-learning module of the branch-and-price-based column generation algorithm. While it effectively identifies rules that capture partial aspects of the ground truth, it occasionally fails to fully match clinical realities. Although stringent, these exact matches hold significant importance in clinical contexts where precision is paramount for diagnosis and treatment. In such scenarios, tailored strategies require more accurate rule-learning capabilities, highlighting the need for improvement in the robustness of our approach.

In real-world applications, the interaction between doctors and AI often occurs within a multi-agent system, where experts

Table 8. Comparison of different models in synthetic experiments without noisy experts. Results are presented as mean ± standard deviation across 10 repeated runs. **Blue shading** indicates the best *independent* model; **Red shading** indicates the best *collaborative* model.

Methods	Accuracy ↑				Rewards ↑	Regret ↓
	Overall	Class-0	Class-1	Class-2		
Independent Models						
AI agent	0.524 ± 0.029	0.468 ± 0.084	0.542 ± 0.077	0.564 ± 0.067	0.652 ± 0.041	0.154 ± 0.065
Doctor1	0.515 ± 0.031	0.562 ± 0.066	**0.585 ± 0.066**	0.397 ± 0.076	0.668 ± 0.055	0.138 ± 0.055
Doctor2	0.470 ± 0.033	0.476 ± 0.078	0.412 ± 0.085	0.521 ± 0.056	0.656 ± 0.028	0.150 ± 0.031
Doctor3	0.529 ± 0.044	0.556 ± 0.080	0.533 ± 0.078	0.497 ± 0.065	**0.676 ± 0.042**	**0.130 ± 0.050**
Doctor4	0.480 ± 0.040	0.559 ± 0.084	0.485 ± 0.054	0.394 ± 0.099	0.617 ± 0.037	0.189 ± 0.048
Doctor5	0.498 ± 0.050	**0.565 ± 0.083**	0.373 ± 0.058	0.555 ± 0.079	0.646 ± 0.025	0.160 ± 0.032
Doctor6	**0.534 ± 0.039**	0.512 ± 0.078	0.512 ± 0.052	**0.579 ± 0.080**	0.669 ± 0.057	0.137 ± 0.086
Doctor7	0.479 ± 0.028	0.391 ± 0.053	0.479 ± 0.063	0.570 ± 0.056	0.642 ± 0.028	0.164 ± 0.052
Collaborative Models						
Majority Voting	0.726 ± 0.038	0.724 ± 0.071	0.712 ± 0.071	0.742 ± 0.061	0.721 ± 0.044	0.085 ± 0.066
Weighted Voting	0.713 ± 0.036	0.726 ± 0.079	0.709 ± 0.080	0.703 ± 0.060	0.740 ± 0.045	0.066 ± 0.051
GLAD	0.271 ± 0.052	0.315 ± 0.079	0.294 ± 0.051	0.203 ± 0.120	0.556 ± 0.054	0.250 ± 0.074
MoE	0.686 ± 0.041	0.768 ± 0.084	0.621 ± 0.084	0.667 ± 0.084	0.709 ± 0.049	0.097 ± 0.049
P+L	0.718 ± 0.035	0.641 ± 0.089	0.752 ± 0.100	0.764 ± 0.052	0.722 ± 0.040	0.084 ± 0.057
SEER	**0.808 ± 0.061**	**0.818 ± 0.104**	**0.839 ± 0.109**	**0.767 ± 0.069**	**0.741 ± 0.044**	**0.065 ± 0.050**
Ablation Study						
SEER(Excluding AI)	0.574 ± 0.053	0.503 ± 0.070	0.597 ± 0.116	0.624 ± 0.062	0.673 ± 0.041	0.133 ± 0.051
SEER(Uncalibrated AI)	0.797 ± 0.041	0.788 ± 0.090	0.812 ± 0.096	0.791 ± 0.055	0.736 ± 0.033	0.070 ± 0.041
P+L(Excluding AI)	0.718 ± 0.035	0.641 ± 0.089	0.752 ± 0.100	0.764 ± 0.052	0.707 ± 0.045	0.099 ± 0.058
P+L(Uncalibrated AI)	0.718 ± 0.035	0.641 ± 0.089	0.752 ± 0.100	0.764 ± 0.052	0.719 ± 0.055	0.087 ± 0.073

may take into account the perspectives of their peers and the AI's recommendations before reaching a final decision. This introduces a more interconnected and complex decision-making environment than our current framework, which treats each expert independently. Furthermore, while our framework supports human intervention to modify or remove high-risk rules, this manual process can be time-consuming and requires substantial expertise. Future work could focus on developing more user-friendly interfaces and automated tools to assist human experts in this task, potentially increasing both efficiency and adoption.

A promising direction for future research involves the introduction of hypernetworks to enable differentiable rule learning, which could improve the accuracy of rule discovery. However, purely data-driven approaches without expert knowledge may introduce noise and fail to capture patient-specific features. Therefore, combining knowledge-based and data-driven frameworks could enhance the robustness and accuracy of rule learning in clinical settings. Moreover, incorporating a human-in-the-loop algorithm would allow for the flexible integration of expert opinions in rule learning, further improving security and stability. While our column-generation algorithm produces relatively stable rules, these enhancements could better align the model with real-world complexities.

M. Computing Infrastructure

All synthetic data experiments are performed on Ubuntu 20.04.3 LTS system with Intel(R) Xeon(R) Gold 6248R CPU @ 3.00GHz, 227 Gigabyte memory.

N. Bayesian Estimation of Confusion Matrix with Partial Feedback

In many real-world scenarios, the true label of an instance is not directly observable. Instead, we receive partial feedback in the form of reward signals. For instance, in medical decision-making, we may only observe whether a patient was accurately diagnosed (e.g., +1) or misdiagnosed (e.g., -1), rather than knowing the exact ground-truth diagnosis. This introduces additional uncertainty in estimating the confusion matrix $\boldsymbol{C}^{(h)}$, which represents the probability that human expert h assigns

Table 9. Comprehensive Rule Set for Diagnosing Gitelman Syndrome. Each rule is assigned a weight representing its relative importance.

#	Rule Description	Weight
	Gitelman Syndrome Rules	
1	$pH > 7.45 \wedge$serum potassium $< 3.0 \wedge$urine potassium $> 20 \wedge$bicarbonate $> 24 \wedge$high blood pressure $= 0$	2.5
2	$pH > 7.45 \wedge$serum potassium $< 3.5 \wedge$urine potassium $> 25 \wedge$bicarbonate $> 24 \wedge$high blood pressure $= 0$	2.5
	Non-Gitelman Syndrome Rules	
3	$pH < 7.35 \wedge$ serum potassium < 3.5	1.5
4	high blood pressure $= 0 \wedge pH < 7.35$	1.3
5	high blood pressure $= 1 \wedge pH > 7.45$	1.3
6	high blood pressure $= 0 \wedge$ urine potassium $< 20 \wedge pH > 7.45 \wedge$ bicarbonate < 22	1.5

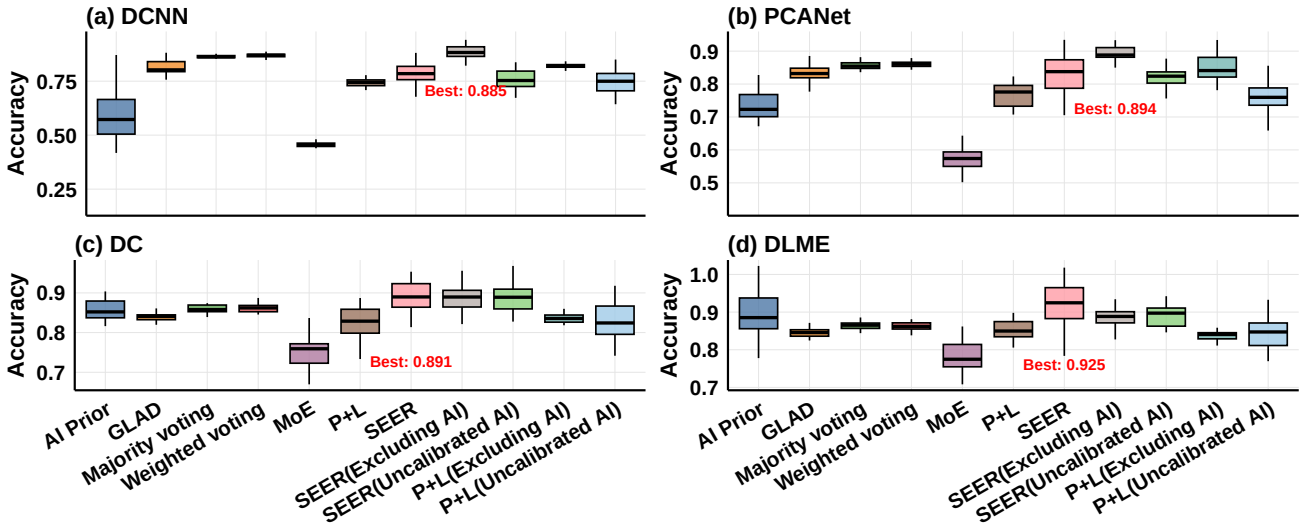

Figure 6. Classification accuracy on CIFAR-10H across base models (10 trials, 200 samples each). Red annotations mark the best methods by mean accuracy.

label j when the true label is i.

To incorporate this feedback into a Bayesian estimation framework, we assume that each row of the confusion matrix follows a Dirichlet prior:

$$\boldsymbol{C}_{i,:}^{(h)} \sim \text{Dirichlet}(\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{iK})$$

where $\alpha_{ij} > 0$ are concentration parameters encoding prior knowledge about the expert's labeling tendencies.

N.1. Incorporating Reward Signals

Since the true label is unknown, we infer it based on the observed reward signals. Let $r_s \in \{+1, -1\}$ denote the reward received for prediction y_s on instance s. We define $P(i \mid r_s, y_s)$, the probability that the true label was i given the observed response y_s and the reward r_s. Using Bayes' rule, this probability is proportional to:

$$P(i \mid r_s, y_s) \propto P(r_s \mid i, y_s) P(i),$$

where:

- $P(r_s \mid i, y_s)$ is the likelihood of receiving reward r_s given that the true label was i and expert h predicted y_s. This

can be modeled using a reward function $f(i, y_s)$, which encodes how likely a correct or incorrect prediction leads to a given reward.

- $P(i)$ is a prior belief about the distribution of true labels.

Given S labeled instances, we update our Dirichlet posterior by marginalizing over possible true labels:

$$\boldsymbol{C}_{i,:}^{(h)} \mid \text{Data} \sim \text{Dirichlet}\left(\alpha_{i1} + \sum_s P(i \mid r_s, y_s)\mathbb{I}(y_s = 1), \ldots, \alpha_{iK} + \sum_s P(i \mid r_s, y_s)\mathbb{I}(y_s = K)\right).$$

This means that the expected count of expert predictions y_s contributing to each row of the confusion matrix is weighted by the inferred probability $P(i \mid r_s, y_s)$, rather than being directly observed.

N.2. MAP Estimation

The Maximum A Posteriori (MAP) estimate for $\boldsymbol{C}_{ij}^{(h)}$ is then given by:

$$\hat{\boldsymbol{C}}_{ij}^{(h)} = \frac{\alpha_{ij} + \mathbb{E}[n_{ij}]}{\sum_{k=1}^{K}(\alpha_{ik} + \mathbb{E}[n_{ik}])},$$

where

$$\mathbb{E}[n_{ij}] = \sum_s P(i \mid r_s, y_s)\mathbb{I}(y_s = j)$$

is the expected count of times label j was assigned when the inferred true label was i.

This formulation enables the estimation of human expert reliability even when we only receive partial reward signals rather than explicit ground-truth labels. By integrating Bayesian inference and probabilistic updates, we construct a robust confusion matrix that reflects both prior knowledge and empirical feedback, helping us model uncertainty in real-world decision-making environments where only outcome-based supervision is available.