Unanchoring the Mind: DAE-Guided Counterfactual Reasoning for Rare Disease Diagnosis

Yuting Yan [1] **Yinghao Fu** [1] **Wendi Ren** [1] **Shuang Li** [1]

Abstract

Diagnosing rare diseases remains a persistent challenge, often hindered by *cognitive anchoring*: once clinicians settle on a common diagnosis, alternative-especially rare-explanations are often dismissed. To address this, we propose a human-centered counterfactual reasoning framework using a Denoising Autoencoder (DAE) to simulate *what-if* diagnostic scenarios that disrupt clinicians' initial assumptions. Our model uniquely jointly learns (1) the true distribution of diseases and symptoms, and (2) human diagnostic behavior, revealing critical gaps between *medically possible* and *clinically considered* diagnoses. By strategically perturbing latent patient representations, it generates *contrastive counterfactuals* that highlight rare-but-plausible conditions—conditions typically overlooked due to cognitive bias. Unlike traditional decision-support tools, our system *proactively* suggests rare diseases not because they are statistically probable, but because they are *cognitively neglected*. Evaluated on three rare disease datasets, our approach outperforms standard machine learning classifiers in detecting rare conditions while maintaining strong performance on common diagnoses. Beyond boosting accuracy, it fosters *hypothesis-driven reasoning*, enhancing both diagnostic precision and clinician learning.

1. Introduction

Despite advances in machine learning for clinical diagnosis, *rare diseases remain notoriously difficult to identify* due to their low prevalence, heterogeneous manifestations, and frequent overlap with more common conditions (Schieppati et al., 2008; Griggs et al., 2009). Consider a patient presenting with persistent fatigue, joint pain, and skin rashes, clinicians often anchor on familiar diagnoses like lupus rather than considering rare alternatives such as Ehlers-Danlos syndrome. This diagnostic misdirection is not merely a result of statistical rarity or symptom ambiguity, but also due to a well-documented *cognitive bias* known as *anchoring*—clinicians' tendency to settle prematurely on an initial diagnosis and insufficiently revise it in light of new or contradictory evidence (Tversky and Kahneman, 1974; Saposnik et al., 2016; Croskerry, 2002; Li et al., 2023).

This *cognitive anchoring* introduces a significant bottleneck in *rare disease detection*, often leading to prolonged diagnostic delays, repeated misdiagnoses, and unnecessary interventions. Studies in clinical cognition have shown that medical decision-making is often driven by fast, heuristic-based thinking that prioritizes pattern recognition over analytical reassessment (Norman et al., 2024). This is especially problematic in the context of rare diseases, where diagnostic presentations often overlap with more common syndromes, creating fertile ground for premature closure. While previous machine learning efforts have primarily focused on enhancing accuracy through larger datasets or more powerful models (Juba and Le, 2019; Sun et al., 2017; Moreno-Barea et al., 2020), few have addressed the cognitive constraints that shape clinicians' interactions with model predictions, particularly under uncertainty. Moreover, existing studies indicate that clinicians may be unable to effectively integrate the AI's reasoning due to its opaque recommendations (Jussupow et al., 2021; Lebovitz et al., 2022), potentially exacerbating misdiagnoses (Jussupow et al., 2022).

Our work tackles the dual challenge of data sparsity and cognitive rigidity by introducing a diagnostic framework that not only *detects rare diseases* but also *mitigates the cognitive biases*—particularly *anchoring*—that hinder accurate diagnosis. Instead of merely maximizing predictive likelihood, our system acts as a cognitive aid, encouraging clinicians to consider alternative diagnostic hypotheses. Drawing from cognitive science theories of bias mitigation (Croskerry, 2002) and leveraging recent advances in generative modeling, we design a Denoising Autoencoder (DAE) (Vincent et al., 2008) generative model to generate plausible

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

Proceedings of the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

diagnostic counterfactuals that promote reflective reasoning.

Our DAE-based model is trained on annotated clinical data to learn both disease distributions and typical diagnostic behaviors. By perturbing the latent representation of a patient's profile, the model generates alternative diagnostic paths—plausible yet cognitively overlooked possibilities—that suggest *follow-up tests*, outside the clinician's immediate expectations. For example, it might suggest:

> *The most likely rare disease overlapping with the current symptoms is **Ehlers-Danlos syndrome**. Consider additional tests such as **genetic screening** for connective tissue disorders. If the results are **positive**, the probability of this diagnosis **increases significantly**.*

Unlike traditional AI systems that deliver static predictions, our framework promotes active cognitive engagement, helping clinicians*break habitual diagnostic patterns* and *rethink their assumptions*. By surfacing rare yet plausible conditions, it expands the diagnostic space, fosters reflective thinking, and supports more informed clinical decisions. As (Buçinca et al., 2021) have demonstrated, a mechanism that guides users to actively engage in critical thinking about initial assumptions enhances decision-making quality more effectively than merely providing predictions.

In our experiments, we evaluate the system's effectiveness using three rare disease datasets. our method outperformed conventional machine learning (ML) classifiers in rare disease detection while preserving optimal performance on common disease diagnosis. Counterfactual *validation* was performed by comparing the model's hypotheses with diagnoses made by *human clinicians* and assessments from *Large Language Models (LLMs)*. The results confirmed that our model could identify plausible but cognitively neglected conditions, thereby enhancing diagnostic precision and fostering clinician learning.

2. Inherent Challenges in Modeling Rare Disease Diagnosis

In clinical diagnosis, the fundamental task is to infer the underlying disease label $Y \in \mathcal{Y}$ from observed clinical evidence $X \in \mathcal{X}$, such as patient-reported symptoms. Both human clinicians and ML models aim to learn or approximate the mapping:

$$h : X \mapsto \hat{Y}, \quad \text{where } \hat{Y} \approx \arg\max_{Y} P(Y \mid X).$$

By Bayes' theorem, this conditional probability can be expressed as:

$$P(Y \mid X) = \frac{P(X \mid Y) \cdot P(Y)}{P(X)},$$

where $P(Y)$ encodes prior knowledge of disease prevalence and $P(X \mid Y)$ reflects the data-generating process (e.g., symptom presentation) conditioned on a specific disease. However, in the context of *rare disease diagnosis*, this inferential process becomes fundamentally challenging, no matter for logistic regression, support vector machines, or even deep classifiers, are all subject to the same three critical limitations:

1. **Skewed priors.** Rare diseases typically have extremely small $P(Y)$. This prior imbalance biases both clinicians and ML models to favor common diagnoses, even when rare diseases are more plausible explanations.

2. **Overlapping symptom profiles.** Many hallmark symptoms of rare diseases (e.g., fatigue, muscle pain, or inflammation) are nonspecific and widely shared across common conditions. As a result, the likelihoods $P(X \mid Y_{\text{rare}})$ and $P(X \mid Y_{\text{common}})$ often overlap significantly, making discrimination between them highly uncertain.

3. **Incomplete evidence.** Key diagnostic features—such as genetic markers or specialized imaging—are frequently missing from the record, due to cost, lack of access, or simply being overlooked. This leads to an underspecified X, causing both humans and machines to rely on incomplete or biased feature sets. Such gaps often *reinforce* cognitive heuristics like *anchoring*, where initial impressions dominate the diagnostic path.

These challenges create a shared *algorithmic–cognitive bottleneck* across both humans and machines. Standard discriminative models $h : X \mapsto Y$, trained to directly map observed features to labels, inherit the same structural vulnerabilities as their human counterparts. Without mechanisms to uncover latent structures, handle missing information, or actively de-bias the inference process, both fall short in the critical task of detecting rare and underrepresented diseases.

2.1. Motivation for a Latent-State Generative Model

These insights motivate the need for a new kind of AI-aided diagnostic framework—one that can:

- *Explicitly identify cases where the observed X lies in an ambiguous or overlapping region* of the feature space;
- *Hypothesize possible latent rare disease explanations* even when current evidence is incomplete;
- *Proactively recommend additional complementary tests* (e.g., genetic panels, imaging) that can disambiguate competing diagnoses and help clinicians *break out of anchored diagnostic pathways*.

A discriminative model alone cannot meet these goals, as it is designed only to map observed input X to a label prediction $\hat{Y}$ and lacks any mechanism for reasoning about uncertainty, missing data, or counterfactual information acquisition. To address these limitations, we propose a *latent-state generative model* based on the Denoising Autoencoder (DAE) framework. This model explicitly learns a latent representation Z of the patient's symptom input X and generates possible reconstructions and diagnostic outcomes in a controlled, interpretable manner. The goal is to assist both machine and human diagnostic reasoning by generating alternative hypotheses—especially those corresponding to rare conditions that might be missed due to low priors or heuristic bias.

The proposed latent-state generative model takes the following form (as illustrated in Fig.1):

- **Input:** X (observed patient symptoms)
- **Latent state:** Z (learned stochastic representation of patient condition)
- **Outputs:**
  1. X': A reconstructed or generated version of patient symptoms (counterfactual or prototypical symptom set)
  2. $\hat{Y}^{\text{AI}}$: Prediction of the true diagnosis based on latent state Z
  3. $\hat{Y}^{\text{human}}$: Model's simulation of a human doctor's likely diagnostic decision

3. Our Proposed Generative Model Formulation

We assume access to a dataset of triplets $\left\{\left(X_i, Y_i^{\text{human}}, Y_i^{\text{true}}\right)\right\}_{i=1}^{N}$, where $X_i \in \mathbb{R}^d$ represents patient features, $Y_i^{\text{true}} \in \{1, \ldots, C\}$ is the ground-truth diagnosis, and Y_i^{human} is the clinician's recorded label. Our goal is to learn a generative latent-state model that captures three components: the patient's latent diagnostic state Z, the clinician's decision Y^{human}, and the AI's prediction Y^{AI}. By explicitly modeling the cognitive gap between human and AI reasoning, the model enables discrepancy-aware inference and supports bias-aware diagnostic support.

$$\begin{aligned} &p_\theta\left(X, Y^{\text{AI}}, Y^{\text{human}}, Z\right) = \\ &p(Z) p_\theta(X \mid Z) p_\theta\left(Y^{\text{AI}} \mid Z\right) p_\theta\left(Y^{\text{human}} \mid \tilde{Z}\right) \end{aligned} \tag{1}$$

Here, $Z \in \mathbb{R}^k$ is a latent representation inferred from X, and $\tilde{Z}$ denotes a modulated version of Z. Although humans and AI observe the same input X, their predictions can diverge due to: (1) cognitive load limiting human attention to parts of X, and (2) fundamentally different mapping functions. We explicitly reflect these factors in the design of our DAE-based generative model.

Latent Representation Learning with Masked Denoising Autoencoder Given that real-world clinical inputs $X \in \mathbb{R}^d$ often contain missing or underreported features, particularly for rare diseases, we employ a masked Denoising Autoencoder (mDAE) (Dupuy et al., 2024) strategy, to learn a robust and informative latent representation Z.

For each observed input X_i, we sample a binary mask $r_i \in \{0,1\}^d$ to randomly drop a subset of observed entries, simulating incomplete or noisy records. The resulting corrupted input is $\tilde{X}_i = r_i \odot X_i$, which is then encoded to a latent distribution $q_\phi\left(Z_i \mid \tilde{X}_i\right)$. The decoder reconstructs the full input, and the reconstruction loss is computed only on the originally observed (i.e., uncorrupted) entries:

$$\mathcal{L}_{\text{recon}} = \mathbb{E}_{q_\phi\left(Z_i \mid \tilde{X}_i\right)}\left[\left\|(1-r_i) \odot \left(X_i - \hat{X}_i\right)\right\|_2^2\right] \tag{2}$$

This approach helps the model infer missing or overlooked features—like masked token prediction in language models—while learning robust, task-relevant representations. These generalizable embeddings enable effective downstream applications such as diagnosis prediction and modeling human-AI divergence.

Dual Classification Losses The latent code Z_i is leveraged to predict two diagnostic outcomes: the *ground-truth diagnosis* Y_i^{true}, and the *observed human diagnosis* Y_i^{human}. We define two separate classification objectives:

- **AI Prediction Loss (truth-matching):**

$$\begin{aligned} &\mathcal{L}_{\text{AI}} = -\mathbb{E}_{q_\phi(Z_i \mid X_i)}\left[\sum_c \alpha_c (1-p_c)^\gamma \log p_c\right] \quad , \\ &\alpha_c \propto \frac{1}{\text{freq}(c)} \end{aligned} \tag{3}$$

Here, $p_c = p_{\theta_{\text{AI}}}\left(Y_i^{\text{true}} = c \mid Z_i\right)$ denotes the predicted probability of class c under the AI classifier. This objective encourages the model to leverage the *full latent representation* Z_i to generate accurate, clinically grounded predictions aligned with the ground-truth diagnosis, using a classifier parameterized by θ_{AI}.

To address class imbalance-particularly prevalent in rare disease settings, we employ a focal loss variant (Lin et al., 2017) that dynamically down-weights well-represented, easily classified categories and emphasizes learning from rare

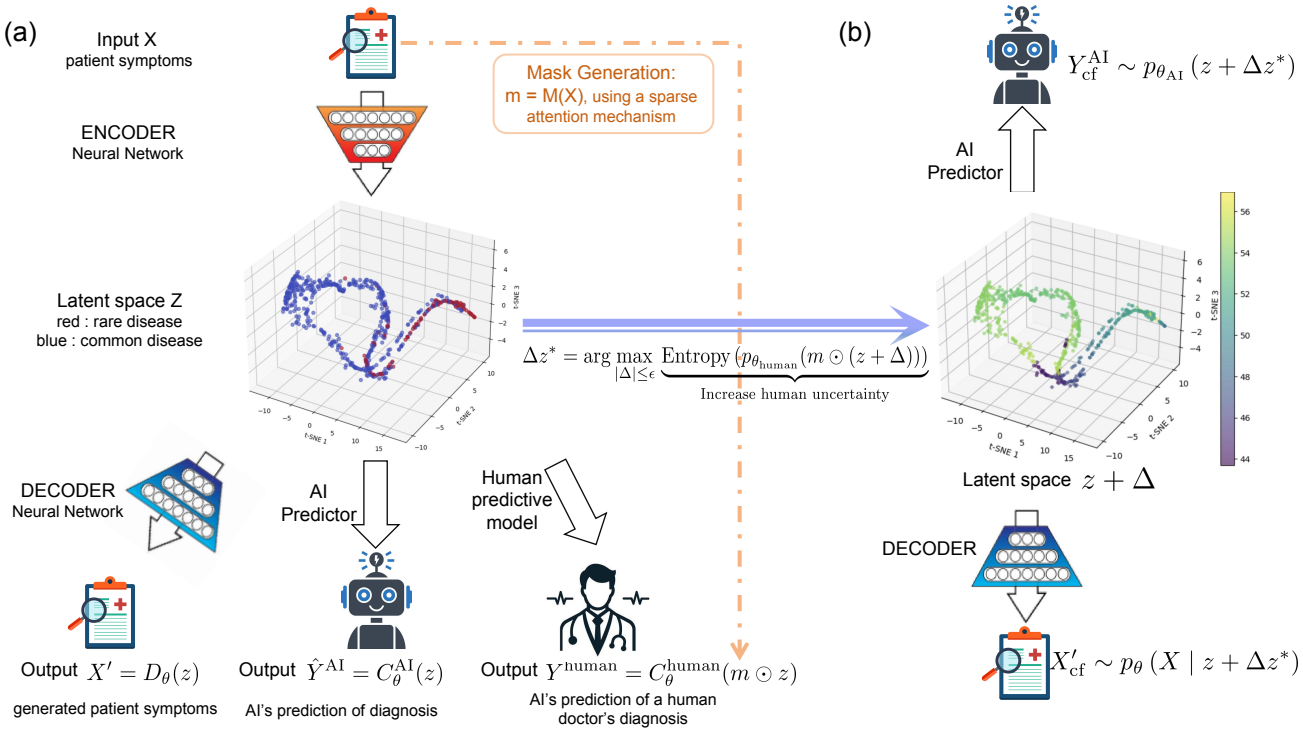


Figure 1. DAE-guided counterfactual reasoning framework. (a) DAE encodes patient features into a latent space, enabling dual predictors for AI and clinician diagnoses. (b) Counterfactuals are generated by perturbing latent vectors along specified directions, with AI producing a list of alternative diagnoses and the decoder creating counterfactual samples that highlight key feature differences, providing clinicians with bias-correcting diagnostic options and showing how slight changes in test results can shift a diagnosis from common disease to rare disease.

or ambiguous cases. As the system is intended to assist clinicians in complex diagnostic scenarios, this calibrated formulation promotes more *exploratory* AI behavior, enabling the model to surface atypical or underrecognized patterns that may otherwise be overlooked. Thus, the AI acts not only as a predictor but also as a discovery aid, supporting more comprehensive and inclusive clinical decision-making.

- **Human Simulation Loss (cognitive-matching):**

$$\mathcal{L}_{\text{human}} = \mathbb{E}_{q_\phi(Z_i|X_i)}\left[-\log p_{\theta_{\text{human}}}\left(Y_i^{\text{human}} \mid \tilde{Z}_i\right)\right] \tag{4}$$

Here, $\tilde{Z}_i = m_i \odot Z_i$ is a selectively masked version of the latent vector, where the learned attention mask $m_i \in [0,1]^k$ gates which latent dimensions are used by the human prediction head. This reflects the idea that, given the same input X_i, *humans and AI may focus on different parts of the data and apply distinct cognitive functions to reach a diagnosis*.

Importantly, the prediction functions for AI and human simulation are parameterized separately, using θ_{AI} and θ_{human} respectively. This architectural asymmetry captures both attentional differences (via m_i) and functional differences in diagnostic reasoning, allowing us to explicitly model and analyze human-AI cognitive divergence.

Modeling Human-AI Cognitive Gaps via Sparse Self-Attention Mask Specifically, we compute the attention mask m_i using a learnable self-attention module:

$$m_i = \text{Softmax}\left(\frac{Q(X_i)\,K(X_i)^\top}{\sqrt{d}}\right)V(X_i) \tag{5}$$

where $Q(\cdot), K(\cdot), V(\cdot)$ are linear projections (as proposed in (Vaswani et al., 2017)) that produce query, key, and value vectors from the input X_i, and the output is pooled to form a k-dimensional attention vector. This attention mechanism identifies which latent features humans are likely to focus on, given the current case.

To ensure interpretability and mimic human cognitive constraints, we impose an ℓ_1 sparsity penalty on the attention mask:

$$\mathcal{L}_{\text{mask}} = \lambda_{\text{mask}} \cdot \|m_i\|_1 \tag{6}$$

This encourages the human prediction head to rely on a small subset of salient features, reflecting *limited cogni-*

tive bandwidth and enhancing the *interpretability* of human diagnostic pathways.

Contrastive Learning for Rare Disease Separability To prevent rare disease embeddings from collapsing into common clusters, we introduce a contrastive loss:

$$\mathcal{L}_{\text{contrast}} = \sum_{(i,j,k)} \max\big(0,\ \delta + d(Z_i, Z_j) - d(Z_i, Z_k)\big),$$

where Z_i and Z_j are latent representations from the same rare disease class, and Z_k is from a common disease class.

This loss encourages embeddings of the same rare class to remain close while pushing them away from embeddings of common classes, thereby promoting greater separability and preserving the distinctiveness of rare conditions in the latent space.

Cognitive Gap Identification: Discrepancy Between AI and Human Attention To quantify the cognitive discrepancy between AI and human reasoning—especially in rare disease cases—we introduce a *cognitive gap loss*. This loss encourages the AI model to attend to features that may be under-utilized by human clinicians, highlighting potential diagnostic blind spots. Formally, we define the loss as:

$$\mathcal{L}_{\text{gap}} = \sum_{i:\, Y_i^{\text{true}} \in \text{rare}} \left\| m_i \odot \nabla_{Z_i} \log p_{\theta_{\text{AI}}}(Y_i^{\text{true}} \mid Z_i) \right\|_2^2,$$

where Z_i is the latent representation, $m_i \in [0,1]^k$ is the learned attention mask approximating human focus, and $\nabla_{Z_i} \log p_{\theta_{\text{AI}}}(Y_i^{\text{true}} \mid Z_i)$ captures the sensitivity of the AI's prediction to each latent feature.

By penalizing high-gradient regions aligned with human attention m_i, the model is encouraged to focus on dimensions that are often overlooked, especially in the context of rare diseases. This fosters attentional divergence in rare disease cases, where the AI can uncover atypical patterns that clinicians might miss due to cognitive biases.

3.1. Total Objective and Training Curriculum

The overall loss function is defined as:

$$\begin{aligned}\mathcal{L}_{\text{total}} =& \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{AI}} + \mathcal{L}_{\text{human}} + \\ & \gamma \mathcal{L}_{\text{contrast}} + \eta \mathcal{L}_{\text{mask}} + \xi \mathcal{L}_{\text{gap}}\,.\end{aligned} \tag{7}$$

The training process follows a staged curriculum, starting with the DAE warm-up using reconstruction loss, followed by the introduction of focal loss for rare disease prediction. The curriculum then adds human cognitive modeling and sparsity regularization, followed by contrastive learning for separating rare and common diseases. Finally, the cognitive gap loss is incorporated to address attention mismatches between Al and human clinicians.

We will train the DAE using the above loss function. Given the learned generative DAE model, we can design the following counterfactual generation tasks.

4. Counterfactual Generation for Cognitive Anchoring Correction

To mitigate diagnostic errors from cognitive anchoring, we introduce a counterfactual generation mechanism that leverages the model's probabilistic structure. Given patient data X, *if $p_{\theta_{\text{AI}}}$ assigns relatively high probability to a plausible diagnosis Y_{AI}—particularly a rare or under-considered one—that diverges from the human's current diagnosis*, this *triggers counterfactual generation* to challenge the initial decision of human and guide follow-up evaluation or testing.

The **goal** of the counterfactual generation here is to

> *Disrupt doctors' fixation on initial hypotheses by generating alternative diagnostic pathways, particularly for rare diseases.*

Learning Optimal Perturbation The perturbation is learned to increase uncertainty in the human (or human-approximating) model, thus exposing cognitive blind spots.

$$\Delta z^* = \arg\max_{\|\Delta\| \le \epsilon} \underbrace{\text{Entropy}\left(p_{\theta_{\text{human}}}(m \odot (z + \Delta))\right)}_{\text{Increase human uncertainty}} \tag{8}$$

Here, $\|\Delta\| \le \epsilon$ ensures that the changes remain within a medically interpretable range. Without perturbation, the Al's prediction from the original z may align closely with the clinician's current belief. By contrast, perturbing z explores latent variations that introduce diagnostic ambiguity from the human's perspective-potentially uncovering under-recognized or rare conditions.

Counterfactual Output Generation Once the optimal perturbation Δz^* is obtained, the system generates two outputs:

- **Al Counterfactual Diagnosis**

$$Y_{\text{cf}}^{\text{AI}} \sim p_{\theta_{\text{AI}}}\left(z + \Delta z^*\right) \tag{9}$$

This may yield a rare disease prediction that prompts reconsideration of the original diagnosis.

- **Synthetic Patient Data Generation** An mDAE is used to reconstruct the corresponding patient profile:

$$X'_{\text{cf}} \sim p_\theta\left(X \mid z + \Delta z^*\right) \tag{10}$$

Here, X'_{cf} represents a plausible synthetic patient who presents similarly but includes key missing symptoms supporting the rare disease.

Finally, the system communicates the counterfactual insight as:

> *"Consider alternative diagnoses with similar presentations: [AI-suggested disease $Y_{\text{cf}}^{\text{AI}}$]. If additional findings such as X'_{cf} were observed, the likelihood of this condition would increase to $p_{\theta_{\text{AI}}}(Y_{\text{cf}}^{\text{AI}} \mid z + \Delta)$."*

This form of explanation aims to encourage the clinician to reflect, reassess, and refine their diagnostic reasoning with evidence-informed support from the AI.

5. Experiment

To evaluate the effectiveness of our proposed framework, we conducted extensive experiments employing three (private) real-world *rare disease* datasets, focused on **Gitelman syndrome**, **acromegaly** and **hypertrophic cardiomyopathy (HCM)**, with detailed specifications provided in Appendix B. These experiments were designed to achieve two primary objectives: first, to validate the robust performance and diagnostic accuracy of our model in rare disease detection; and second, to evaluate the efficacy of counterfactuals in addressing cognitive gaps and guiding clinical decision-making.

5.1. Model Performance

5.1.1. Prediction on Imbalanced Data

The low prevalence of rare diseases inherently leads to imbalanced datasets. This disparity poses a significant challenge for conventional machine learning classifiers, which are often sensitive to such imbalances. We conducted experiments on the real-world Gitelman dataset, where the imbalance ratio (the number of common disease samples divided by the number of rare disease samples) varies from 94:100 to 94:500. Our approach outperforms five typical machine learning classifiers, as illustrated in Fig. 2 (from left to right, the figure shows AUC, ACC for common diseases, and ACC for rare diseases). Notably, our model's AUC increases as the imbalance ratio grows because the larger overall data volume provides more information for learning despite the greater skew. These results demonstrate our method's robustness and reliability in capturing meaningful patterns in increasingly imbalanced data.

5.1.2. Latent Space Visualization

We visualize the model's latent space using the Gitelman dataset in three distinct ways. These visualizations, shown in Fig. 3, offer valuable insights into the model's internal representations. Panel (a) shows the structural organization of latent embeddings, illustrating the model's ability to encode fine-grained phenotypic details that distinguish clinically similar samples. Panel (b) presents an attention map of clinician focus within the same space: mask values of 1 mark high clinical relevance regions, while 0 indicates lower priority, directly aligning attention with diagnostic importance. Panel (c) highlights features exerting significant influence on human classification decisions, exposing potential decision boundaries where predictions may shift. The visualization principle involves perturbing latent space vectors to maximize human prediction uncertainty, with the intensity distribution directly reflecting perturbation magnitude. Lighter colors denote higher diagnostic uncertainty, revealing critical knowledge gaps that could lead to misdiagnosis.

Additionally, we conducted an ablation study to evaluate the necessity of each loss term in our model's total loss function. As shown in Appendix C , our findings indicate that removing any single loss term negatively impacts the model's performance.

5.2. Counterfactual Sample Generation

Addressing the pain points and diagnostic needs in rare disease medicine, our model enables counterfactual analysis experiments across diverse scenarios. Here, we focus on three typical and practical scenarios for detailed evaluation:

Scenario 1: Feature Completion for Low-Confidence Predictions: When a patient's original input features have missing values, overlap significantly with common disease characteristics, and yield low-confidence AI predictions for common diagnoses, our model generates counterfactual samples to address missing features. This refines clinical judgments and guides decision-making.

Scenario 2: AI-Human Prediction Discrepancy Resolution: When discrepancies exist between AI predictions and clinician diagnoses, our model generates flipped samples to reveal differences in decision-making logic. This provides clinicians with interpretable insights to reconcile divergent conclusions.

Scenario 3: Uncertainty-Driven Alternative Diagnoses: By perturbing feature vectors in latent spaces where clinicians exhibit maximal diagnostic uncertainty, our model generates alternative diagnosis lists. This anchors cognitive bias correction and supports robust differential diagnosis.

5.3. Qualitative Evaluation of Counterfactuals

For a more comprehensive assessment, an LLM- and doctor-based evaluation framework is designed for evaluating counterfactual outcomes.

Evaluated by LLM Since 2023, LLMs with advanced instruction-following and semantic comprehension have en-

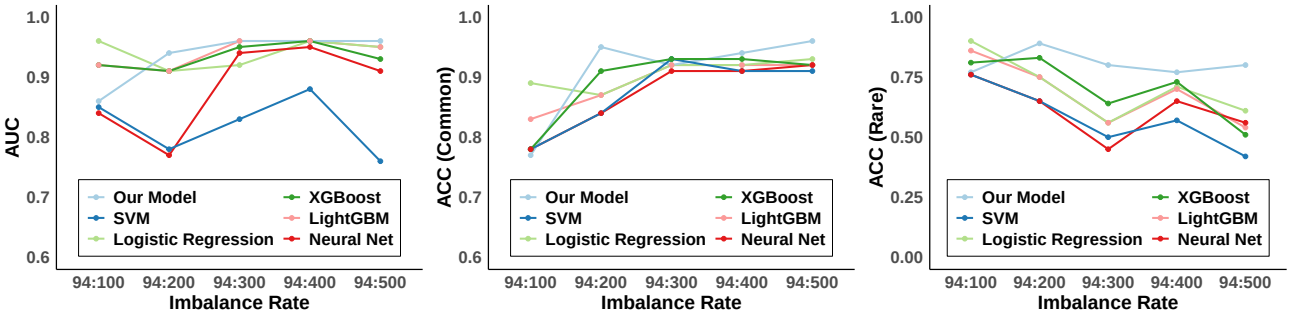
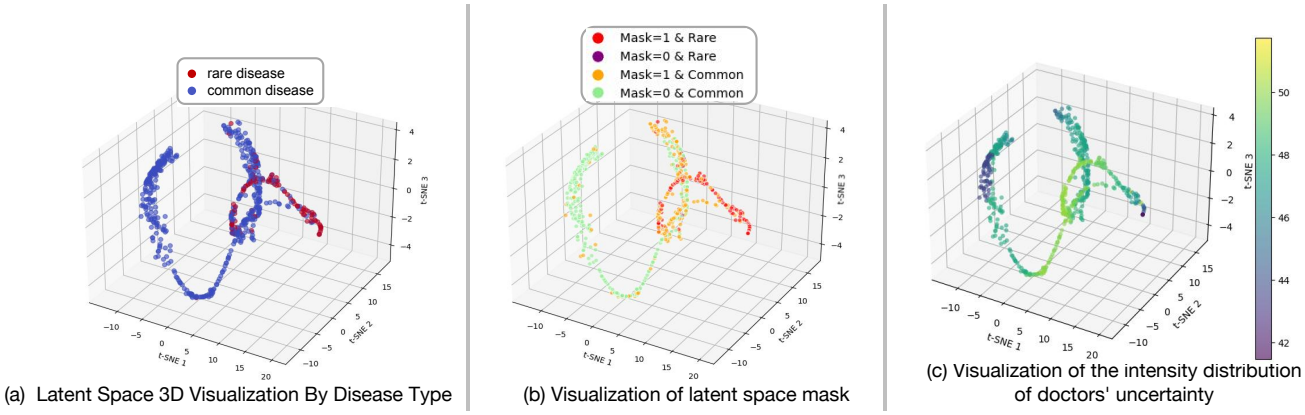


Figure 2. Comparison of model performance under imbalanced data.



(a) Latent Space 3D Visualization By Disease Type

(b) Visualization of latent space mask

(c) Visualization of the intensity distribution of doctors' uncertainty

Figure 3. Latent space visualization by disease type, clinician attention, and diagnostic uncertainty.

abled automated evaluation (Gao et al., 2025). In our framework, pre-trained LLMs assess generated counterfactual instances via structured prompts emulating human judgment, evaluating criteria like plausibility, clinical relevance and cognitive enhancement potential. LLMs systematically quantify counterfactual effectiveness across semantic, causal, and operational dimensions.

Evaluated by Doctors Clinical experts from a leading hospital validated rare disease counterfactuals for medical plausibility and clinical relevance, leveraging their domain expertise.

Fig. 4 shows Prompt, LLM evaluations and doctor evaluations across three scenarios, including case segments from three datasets and core elements of the LLM prompting framework (roles, instructions, evaluation forms). Empirical analyses and prompting details are integrated to demonstrate diagnostic evaluation structures

For LLM prompting specifics and responses, See Appendix D.

5.4. Quantitative Evaluation of Counterfactuals against Baselines

In addition to the above qualitative approach, We also conducted quantitative comparative experiments, where we consider two baseline methods for generating counterfactuals. This dual quantitative-qualitative approach ensures comprehensive validation of the scientific validity and practical utility of counterfactual samples in advancing rare disease research.

5.4.1. Experimental Setup

Baselines We consider two baseline methods for generating counterfactuals: REVISE (Joshi et al., 2019) and CF-VAE (Nagesh et al., 2023). REVISE employs an optimization-based strategy within the latent space of a generative model to produce counterfactuals. In contrast, CF-VAE, a counterfactual variational autoencoder, does not rely on latent space optimization; instead, it jointly optimizes the variational autoencoder alongside a subsequent binary prediction model.

Metrics We quantitatively evaluate the counterfactuals generated by our method and baseline models across two

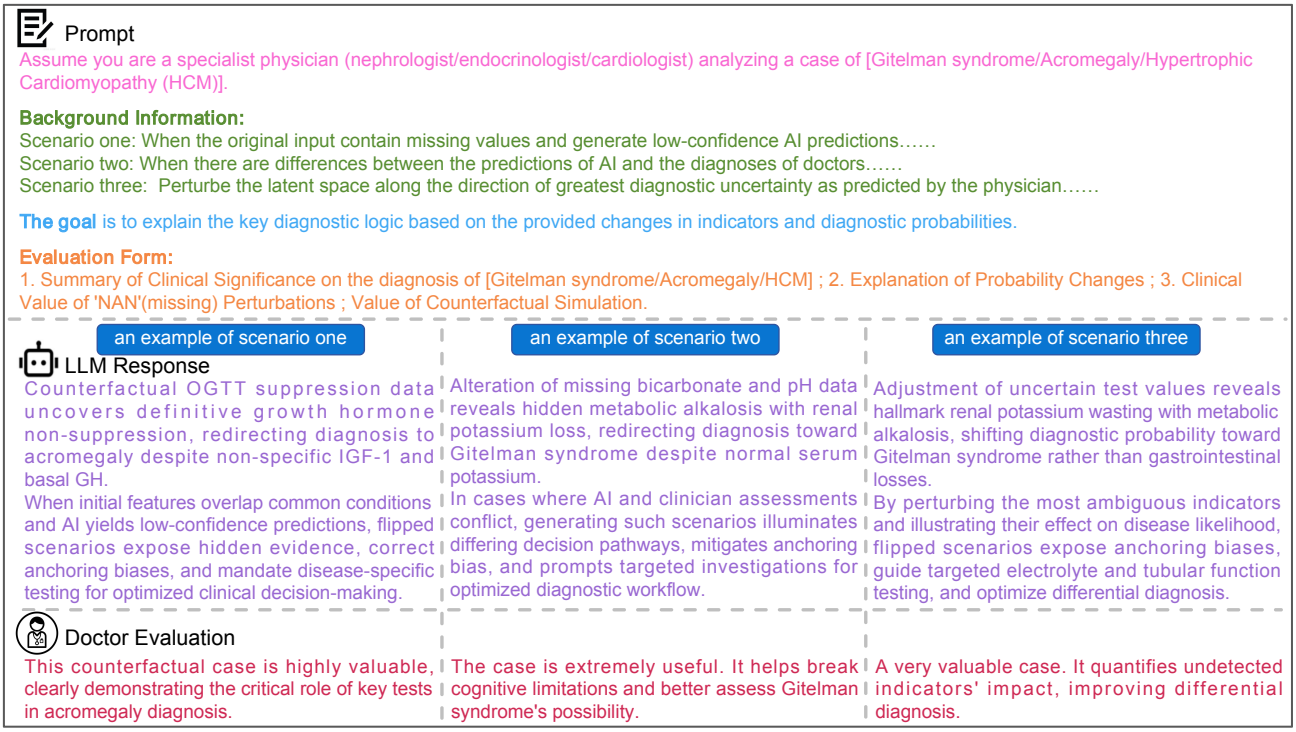


Figure 4. Illustration of prompt, LLM response segment and real world doctor evaluation segment.

Table 1. Performance metrics across three datasets.

Dataset	Model	Label Flip Rate	RMSE
Gitelman	REVISE	0.96±0.03	5.40±0.89
	CFVAE	0.96±0.02	12.00±1.77
	Our Model	**1.00±0.00**	**1.93±0.76**
	Ablation	**1.00±0.00**	4.85±3.27
Acromegaly	REVISE	0.92±0.11	13.96±14.44
	CFVAE	0.85±0.15	13.96±14.84
	Our Model	**1.00±0.00**	**0.18±0.10**
	Ablation	**1.00±0.00**	0.21±0.08
HCM	REVISE	0.70±0.40	0.33±0.04
	CFVAE	0.80±0.40	0.33±0.01
	Our Model	**1.00±0.00**	**0.10±0.13**
	Ablation	**1.00±0.00**	0.46±0.27

key dimensions: (1) **Label Flip Rate**: For the binary classification tasks across our three real-world datasets, label flip rate measures the proportion of counterfactuals classified by the target model into the intended target class, assessing their validity. (2) **Root Mean Squared Error (RMSE)**: RMSE is computed between each generated counterfactual and its corresponding original input to quantify the perturbation magnitude with a lower RMSE suggesting higher plausibility.

5.4.2. Comparisons with Baseline Methods

Table 1 compares our model, two baseline methods(REVISE and CFVAE) and an ablation experiment across three datasets. The ablation experiment removes the final multi-loss fine-tuning phase, retaining only stage-wise training of the DAE, AI predictor, human predictor, and mask network. Our model achieves the highest Label Flip Rate and lowest RMSE across all datasets, indicating superior effectiveness in generating relevant counterfactuals with minimal perturbation.

6. Conclusion

We introduced a human-centered counterfactual reasoning framework that perturbs latent patient representations via a DAE-based latent state generative model to counteract cognitive anchoring in rare disease diagnosis. By generating realistic and clinically plausible "what-if" scenarios, our method surfaces overlooked conditions and systematically guides clinicians toward alternative diagnostic hypotheses. A comprehensive mixed evaluation involving both large language models (LLMs) and medical professionals confirms the scientific soundness, clinical relevance, and interpretive clarity of the generated cases. This framework fosters reflective diagnostic reasoning, enhances interpretability, and offers a scalable tool for bridging human knowledge gaps in challenging medical scenarios.

Impact Statement

This work introduces a counterfactual reasoning framework to address cognitive biases in rare disease diagnosis, potentially reducing diagnostic delays and improving patient outcomes by prompting clinicians to consider overlooked conditions. The approach prioritizes ethical considerations through data anonymization and interpretable AI design, ensuring it supplements rather than replaces clinical judgment. By modeling human-AI cognitive gaps, the framework advances responsible AI in healthcare, with broader implications for mitigating biases in high-stakes decision-making domains.

References

Arrigo Schieppati, Jan-Inge Henter, Erica Daina, and Anita Aperia. Why rare diseases are an important medical and social issue. *The Lancet*, 371(9629):2039–2041, 2008.

Robert C Griggs, Mark Batshaw, Mary Dunkle, Rashmi Gopal-Srivastava, Edward Kaye, Jeffrey Krischer, Tan Nguyen, Kathleen Paulus, Peter A Merkel, et al. Clinical research for rare disease: opportunities, challenges, and solutions. *Molecular genetics and metabolism*, 96(1): 20–26, 2009.

Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131, 1974.

Gustavo Saposnik, Donald Redelmeier, Christian C Ruff, and Philippe N Tobler. Cognitive biases associated with medical decisions: a systematic review. *BMC medical informatics and decision making*, 16:1–14, 2016.

Pat Croskerry. Achieving quality in clinical decision making: cognitive strategies and detection of bias. *Academic emergency medicine*, 9(11):1184–1204, 2002.

Alexander C Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2206–2217, 2023.

Geoff Norman, Thierry Pelaccia, Peter Wyer, and Jonathan Sherbino. Dual process models of clinical reasoning: the central role of knowledge in diagnostic expertise. *Journal of Evaluation in Clinical Practice*, 30(5):788–796, 2024.

Brendan Juba and Hai S Le. Precision-recall versus accuracy and the role of large data sets. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 4039–4048, 2019.

Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017.

Francisco J Moreno-Barea, José M Jerez, and Leonardo Franco. Improving classification accuracy using data augmentation on small data sets. *Expert Systems with Applications*, 161:113696, 2020.

Ekaterina Jussupow, Kai Spohrer, Armin Heinzl, and Joshua Gawlitza. Augmenting medical diagnosis decisions? an investigation into physicians' decision-making process with artificial intelligence. *Information Systems Research*, 32(3):713–735, 2021.

Sarah Lebovitz, Hila Lifshitz-Assaf, and Natalia Levina. To engage or not to engage with ai for critical judgments: How professionals deal with opacity when using ai for medical diagnosis. *Organization science*, 33(1):126–148, 2022.

Ekaterina Jussupow, Kai Spohrer, and Armin Heinzl. Radiologists' usage of diagnostic ai systems: The role of diagnostic self-efficacy for sensemaking from confirmation and disconfirmation. *Business & Information Systems Engineering*, 64(3):293–309, 2022.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008.

Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-computer Interaction*, 5(CSCW1):1–21, 2021.

Mariette Dupuy, Marie Chavent, and Remi Dubois. mdae: modified denoising autoencoder for missing data imputation. *arXiv preprint arXiv:2411.12847*, 2024.

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007, 2017.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Neural Information Processing Systems*, 2017.

Mingqi Gao, Xinyu Hu, Xunjian Yin, Jie Ruan, Xiao Pu, and Xiaojun Wan. Llm-based nlg evaluation: Current status and challenges. *Computational Linguistics*, pages 1–28, 2025.

Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. Towards realistic individual recourse and actionable explanations in black-box decision making systems. *arXiv preprint arXiv:1907.09615*, 2019.

Supriya Nagesh, Nina Mishra, Yonatan Naamad, James M Rehg, Mehul A Shah, and Alexei Wagner. Explaining a machine learning decision to physicians via counterfactuals. In *Conference on Health, Inference, and Learning*, pages 556–577. PMLR, 2023.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31: 841, 2017.

Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 607–617, 2020.

Sahil Verma, John Dickerson, and Keegan Hines. Counterfactual explanations for machine learning: A review. *arXiv preprint arXiv:2010.10596*, 2(1):1, 2020.

Min Hun Lee and Chong Jun Chew. Understanding the effect of counterfactual explanations on trust and reliance on ai for human-ai collaborative clinical decision making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–22, 2023.

Eleni Straitouri, Suhas Thejaswi, and Manuel Rodriguez. Controlling counterfactual harm in decision support systems based on prediction sets. *Advances in Neural Information Processing Systems*, 37:129443–129479, 2024.

Arnaud Van Looveren and Janis Klaise. Interpretable counterfactual explanations guided by prototypes. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 650–665. Springer, 2021.

Daniel Nemirovsky, Nicolas Thiebaut, Ye Xu, and Abhishek Gupta. Countergan: Generating counterfactuals for real-time recourse and interpretability using residual gans. In *Uncertainty in Artificial Intelligence*, pages 1488–1497. PMLR, 2022.

Stig Hellemans, Andres Algaba, Sam Verboven, and Vincent Ginis. Flexible counterfactual explanations with generative models. *arXiv preprint arXiv:2502.17613*, 2025.

Wenzhuo Yang, Jia Li, Caiming Xiong, and Steven CH Hoi. Mace: An efficient model-agnostic framework for counterfactual explanation. *arXiv preprint arXiv:2205.15540*, 2022.

Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. *Advances in neural information processing systems*, 32, 2019.

A. Related work

Counterfactual Explanations The evolution of counterfactual explanations has transitioned from optimizing feature perturbations (Wachter et al., 2017) to frameworks that prioritize human-AI collaboration and safety. Early methods focused on generating minimal feasible changes (e.g., DiCE (Mothilal et al., 2020)), but were criticized for ignoring user-specific constraints and real-world applicability (Verma et al., 2020). More recent work, including (Lee and Chew, 2023), highlights the role of counterfactuals in mitigating cognitive biases. (Lee and Chew, 2023) showed that exposing users to hypothetical scenarios reduces overreliance on erroneous AI predictions, particularly among non-experts susceptible to confirmation bias. This aligns with broader findings in human-AI interaction, where explanations must balance interpretability with decision accuracy (Buçinca et al., 2021; Straitouri et al., 2024). A significant advancement in this area is the formalization of counterfactual harm, defined as the risk that explanations may degrade human judgment. (Straitouri et al., 2024) introduced structural causal models with conformal risk control to bound harmful outcomes in clinical systems. Their approach integrates monotonicity assumptions (e.g., "higher biomarker values correlate with worse prognosis") to ensure explanations align with domain knowledge, thereby addressing a gap in earlier optimization-based methods (Van Looveren and Klaise, 2021). This shift reflects a growing emphasis on safety-critical metrics, moving beyond traditional criteria like sparsity and realism (Verma et al., 2020).

Counterfactual Generative Models Generative models have been introduced to generate numerical counterfactuals, enabling dynamic adaptation to user constraints. Early GAN-based approaches, such as CounterRGAN (Nemirovsky et al., 2022), enforced immutable features via residual networks but lacked flexibility for real-time customization. FCEGAN (Hellemans et al., 2025) addresses this limitation by incorporating user-defined templates and dual discriminator losses, facilitating personalized explanations in domains like loan approvals (Yang et al., 2022). These frameworks align with CTGAN's training-by-sampling strategy (Xu et al., 2019) to handle class imbalance, a persistent challenge in financial and medical datasets. While REVISE (Joshi et al., 2019) introduced a method for generating numerical counterfactuals using arbitrary generative models, it can produce unrealistic counterfactuals, making them unsuitable for healthcare applications, and is limited by the need for multiple calls to an optimization module. Although CFVAE (Nagesh et al., 2023) was designed for generating counterfactuals in healthcare settings using variational autoencoders, it does not account for realistic challenges in healthcare, such as class imbalance in rare disease cases and missing values in datasets. To overcome these limitations, we propose a novel method designed for healthcare applications, particularly in rare disease diagnosis. Our approach generates personalized counterfactuals for clinicians while handling missing values and class imbalance in the training data.

B. Experimental Datasets

To evaluate our method, we consider the following three private datasets.

Gitelman Syndrome This dataset comprises real clinical records from a top hospital, focusing on Gitelman syndrome (GS), a rare autosomal recessive renal tubulopathy. The data contains 594 patients, including 94 diagnosed with GS and 500 non-GS individuals. Five key diagnostic features are included: *Serum Potassium*, *Urine Potassium*, *pH*, *Bicarbonate*, and *High Blood Pressure*, with labels derived from clinical diagnoses. To emulate real-world scenarios where critical test results are missing (a common challenge in rare disease diagnosis), we artificially mask a subset of these features in the original data (initially complete) by replacing values with NaN. This enables counterfactual analysis to quantify how missing tests impact predictions, thereby guiding clinicians to prioritize specific examinations for undiagnosed cases. The dataset is split into 80%-20% train-test sets for GS classification, with subsequent counterfactual perturbation analysis performed in the latent space of the complete data. It should be noted that we retained the situation of data imbalance, which is to be consistent with the situation that the incidence of rare diseases in the real world is much lower. And despite this imbalance, our model still maintained good performance.

Acromegaly This dataset includes real-world clinical records from a top hospital, focusing on acromegaly, a chronic disorder caused by excessive growth hormone (GH) secretion, typically due to pituitary somatotroph adenomas. The data contains 181 patients, comprising 88 diagnosed with acromegaly and 93 non-acromegaly controls. Three clinically significant features are incorporated: *Serum GH*, *IGH-1*, and *OGTT-GH_min*, with labels derived from clinical diagnoses. To reflect realistic data incompleteness, we retain naturally occurring missing values in the original dataset and explicitly record their positions. This facilitates counterfactual generation that aligns with clinical practice, allowing clinicians to evaluate how incomplete laboratory profiles influence diagnostic predictions. The dataset is partitioned into 80%-20% training-test sets for binary

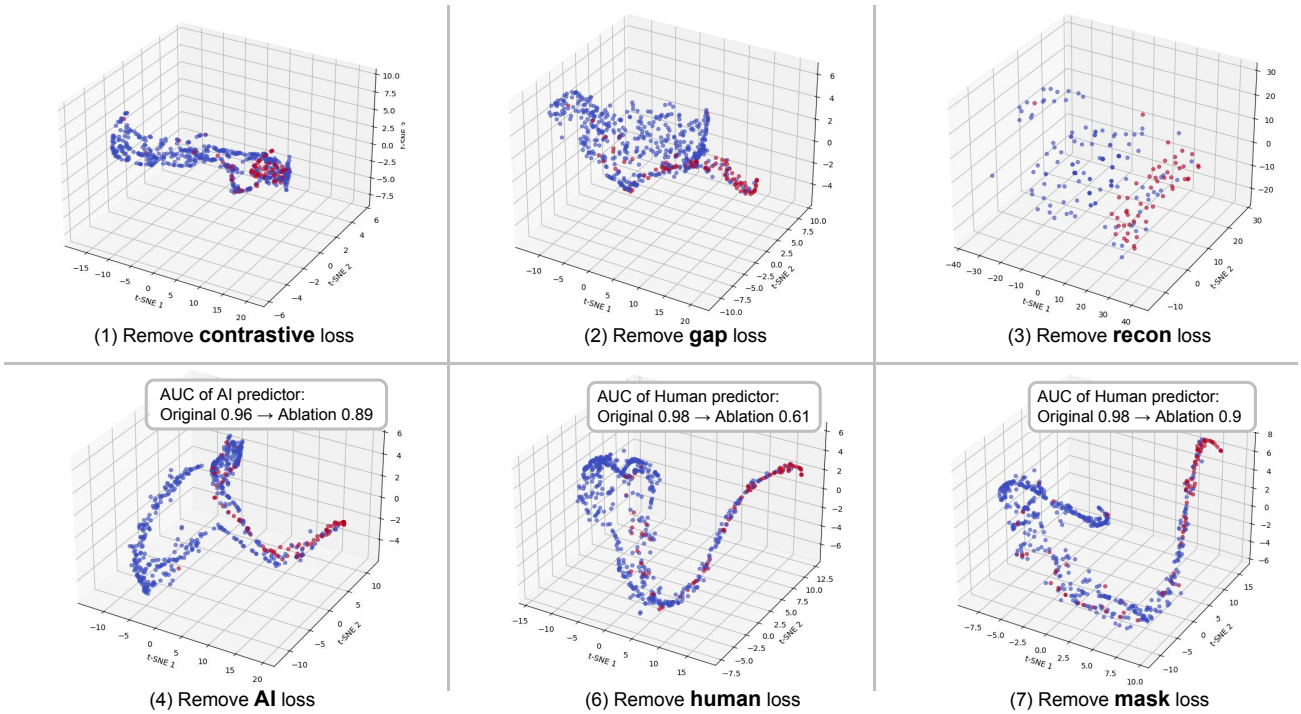


Figure 5. Ablation study: loss function removal impact on latent space and model performance.

classification, followed by counterfactual perturbation and interpretability analysis in the latent space of the complete data to identify critical diagnostic drivers.

Hypertrophic Cardiomyopathy (HCM) This dataset includes real-world clinical records from a top hospital, focusing on hypertrophic cardiomyopathy (HCM), an inherited cardiac disorder characterized by abnormal myocardial thickening that may lead to ventricular outflow tract obstruction, arrhythmias, and heart failure. The data contains 36 patients, including 21 HCM-diagnosed individuals and 15 individuals with another rare disease (ATTR, amyloidosis trans-thyretin related) as the control group. Eight clinically significant features are incorporated: *Asymmetric Hypertrophy*, *SAM*, *Low Left Ventricular Voltage*, *High Left Ventricular Voltage*, *Family History*, *Sarcomere Gene Mutation*, *TTR Gene Mutation*, and *Amyloid Deposition*. Similarly, to preserve clinical authenticity, naturally occurring missing values in the original dataset are retained and explicitly mapped for interpretability. The dataset is partitioned into 80%-20% training-test splits for HCM classification. Post-training, counterfactual perturbation and causal analysis are conducted in the latent space of the complete data to identify critical diagnostic patterns and feature interactions.

C. Latent Space Visualization with Ablation Study

We conduct an ablation study to evaluate the necessity of each loss term in our model's total loss function. Specifically, we visualize the distribution of the latent space when individually removing each loss component during fine-tuning (prior to fine-tuning, each component of our model, including DAE, AI predictor, mask net and human predictor, is first trained in stages with its corresponding loss function). As shown in Fig. 5, Our findings indicate that the removal of the contrastive loss, gap loss, or reconstruction loss degrades the quality of the latent space representation, thereby impairing the model's ability to discriminate between similar samples. In contrast, removal of the AI prediction loss, AI prediction loss or mask regularization loss impairs the performance of the AI predictor or human predictor, as depicted by the AUC changes in the figure, underscoring the indispensable role of each loss component in maintaining model effectiveness.

D. Details of Prompting LLM and Counterfactual Evaluations

Fig. 6 illustrates the operational mechanism of prompting the LLM and LLM response across three counterfactual scenarios. For each scenario, a representative case is selected: the first from the acromegaly dataset, and the latter two from the Gitelman dataset. This visual depiction not only offers profound insights into the framework's functionality but also provides a practical reference for clinicians and researchers, underscoring the significance of counterfactual reasoning in enhancing the differential diagnosis of rare diseases.

E. Model Architecture Details

E.1. DAE Architectures

The Denoising Autoencoder (DAE) architecture captures clinical feature mappings through an Encoder and Decoder. The Encoder uses ELU activations to project raw features into a 32-dimensional latent space, while the Decoder reconstructs inputs from this space. Categorical features are embedded via a dedicated layer, and the design supports robust learning from incomplete data. Take the Gitelman syndrome dataset as an example, key components are detailed in Table 2, which outlines layer dimensions and functional roles.

Table 2. DAE architecture configuration

Component	Layers	Dimension	Functional Description
Encoder	Input Layer	5	Raw clinical features
	Hidden Layer	128	ELU-activated transformation: $h = \text{ELU}(Wx + b)$
	Latent Space	32	Bottleneck representation: z
	Embedding	8	Categorical feature encoding: $\text{onehot}(x)W_e$
Decoder	Input Layer	32	Latent space input: z
	Hidden Layer	128	Feature decoding: $h_d = \text{ELU}(W_d z + b_d)$
	Output Layer	5	Feature reconstruction: $\hat{x}$

E.2. Predictor Architectures

The AI and human predictors, along with the attention mask network, are designed to explicitly model the divergence between machine and clinician reasoning. The AI predictor operates in the full latent space to generate ground truth-aligned diagnoses, while the human predictor uses a sparse attention mask (generated by the mask network) to simulate cognitive constraints in clinical decision making. Table 3 outlines the architecture details, including layer dimensions, activation functions, and the attention mechanisms. This modular design supports interpretable counterfactual generation by isolating human-AI cognitive gaps in the latent space.

Table 3. Predictor Architectures Configuration

Component	Layers	Dim/Num of Heads	Description
AI Predictor	Input Layer	32	ELU-activated projection into hidden space
	Hidden Layer	128	ELU transformation of latent features
	Output Layer	2	Produces class logits for prediction
Mask Network	Input Layer	5	ELU-activated linear embedding
	Attention Layer	4	Multi-head self-attention for contextual feature interaction
	Output Layer	32	Generates masking coefficients
Human Predictor	Input Layer	32	Takes the masked latent representation as input
	Hidden Layer	128	ELU transformation of masked latent space
	Output Layer	2	Produces class logits aligned with experts

F. Training Configuration Details

F.1. Stage-Wise Training Details

The model is trained in four stages: DAE warm-up, AI predictor training, joint human predictor and mask network training, and fine-tuning. Table 4 specifies the learning rate schedules, batch sizes, and regularization strategies (e.g., gradient clipping) for each phase on the Gitelman syndrome dataset. For instance, the DAE warm-up phase employs learning rate annealing and early stopping to stabilize latent space initialization. This staged approach balances model complexity and training stability while ensuring task-specific optimization.

Table 4. Progressive training strategy

Phase	Components	Learning Rate	Key Details
DAE Train	Encoder / Decoder	1e-4	• LR annealing • Early stop • Gradient clip ≤ 1.0 • Batch size 16
AI Predictor Train	AI Predictor Network	1e-4	• LR annealing • Early stop • Gradient clip ≤ 1.0 • Batch size 16
Human Predictor + Mask Net Train	Human Predictor Network, Mask Network	1e-4	• LR annealing • Early stop • Gradient clip ≤ 1.0 • Batch size 16
Fine-Tuning	Full Network	1e-4	• Gradient clip ≤ 1.0 • Batch size 16

F.2. Loss Function Weight in Fine-Tuning Stage

The total training loss combines multiple objectives, including reconstruction, classification, contrastive separation, and cognitive gap minimization. Table 5 defines the weights assigned to each loss component on the Gitelman syndrome dataset, emphasizing the balance between feature reconstruction (dominant in early stages) and rare/common disease separability (enforced via contrastive loss).

Table 5. Loss Function Specification

Loss Type	Weight	Function
Reconstruction	1	Reconstruct input features
AI	1	Maximize AI prediction accuracy
Human	1	Align with human diagnoses
Mask	0.001	Promote sparse attention masks masks
Contrastive	1.5	Separate rare/common diseases
Gap	1.5	Reduce human-AI attention gaps

G. Broader Impact and Limitation

This study aims to address the underdiagnosis of rare diseases caused by cognitive biases in clinical decision-making. Our framework helps clinicians consider rare conditions more effectively through generative counterfactuals, potentially reducing diagnostic delays and improving patient outcomes, especially in underserved areas with limited specialized expertise. By modeling the cognitive gaps between humans and AI, it promotes transparent and bias-aware collaboration, setting a practical example for AI applications in healthcare and other high-stakes fields.

Potential risks include the possibility of over-relying on AI, which we mitigate by designing interpretable counterfactual explanations to supplement, rather than replace, clinical judgment. Data privacy and bias issues are addressed through strict data anonymization and cross-population validation to ensure fairness. Ethically, the research focuses on diagnostic support

rather than treatment decisions and uses synthetic data for validation to minimize risks.

However, While our framework demonstrates promise, several limitations warrant consideration. First, the datasets are derived from specialized hospitals, which may limit generalizability to diverse healthcare settings or underrepresented populations. Second, while validated on three rare diseases, the model's effectiveness on ultra-rare conditions remains untested. Finally, the computational cost of latent space perturbation may hinder real-time deployment in resource-constrained environments.

In conclusion, this work balances technical innovation with ethical considerations, providing a scalable tool to advance rare disease diagnosis and foster collaborative AI-driven healthcare. Its approach offers valuable insights for AI-assisted decision-making in complex domains.

H. 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.

System Prompt	Assume you are a specialist physician (nephrologist/endocrinologist/cardiologist) analyzing a case of [Gitelman syndrome/Acromegaly/Hypertrophic Cardiomyopathy (HCM)]. Background Information: The counterfactual changes in clinical indicators in the following case are generated by perturbing the model along the direction of greatest diagnostic uncertainty as predicted by the physician. This method aims to provide a data-driven alternative perspective that may differ from the initial clinical judgment, helping to correct cognitive anchoring and enabling a more comprehensive assessment of rare diseases. The goal is to explain the key diagnostic logic based on the provided changes in indicators and diagnostic probabilities. Important Note for HCM: The HCM-related indicators (e.g., asymmetric hypertrophy, left ventricular voltage, family history, etc.) are binary variables (0 or 1), where 0 typically indicates negative/normal and 1 indicates positive/abnormal. These are not continuous physiological measurements. The final evaluation should include: Summary of Clinical Significance: Summarize the overall impact of key indicator changes on the diagnosis of [Gitelman syndrome/Acromegaly/HCM]. Explanation of Probability Changes: Summarize the main reasons for the changes in diagnostic probabilities. Clinical Value of 'nan' Perturbations (if applicable): Summarize the significance of perturbing 'nan' (missing) values for prompting further tests and diagnosing rare diseases. Value of Counterfactual Simulation: Briefly describe, based on its generation mechanism (perturbing in the direction of greatest uncertainty to correct cognitive anchoring), how this simulation helps clinical diagnosis, especially in avoiding premature exclusion of rare diseases.		
User Prompt	The key indicator changes for case {i+1} are as follows, where the original value of 'nan' indicates that the test was not performed: [Disease-specific indicators, e.g., for Gitelman: Urine Potassium, Bicarbonate, Serum Potassium, High Blood Pressure, pH; for Acromegaly: IGF-1, Serum GH, OGTT-GH; for HCM: Asymmetric Hypertrophy, Low/High Left Ventricular Voltage, Family History, Amyloid Deposition, Sarcomere Gene Mutation, LVOTO, SAM, TTR Gene Mutation] Diagnostic probability changes: Original (Common Disease/Rare Disease): {p_common}/{p_rare} After Perturbation (Common Disease/Rare Disease): {p_common_perturb}/{p_rare_perturb} Please provide a detailed process analysis and result evaluation based on the above data and your medical knowledge.		
LLM Response	Case 1: Initially relying solely on IGF-1 and basal GH levels may not clarify the diagnosis (both may be at critical values or nonspecific), leading to similar probabilities for the rare disease (acromegaly) and common diseases (e.g., physiological GH fluctuations, other metabolic disorders). After perturbation, the OGTT-GH suppression test result directly confirms the diagnosis. The failure of GH to suppress during OGTT—a key pathological feature of acromegaly—exhibits extremely high specificity, effectively ruling out other common diseases. Omission of the critical OGTT-GH suppression test introduces diagnostic uncertainty and may lead to missed acromegaly (which requires OGTT for confirmation). The perturbed scenario highlights the irreplaceable role of this test. This simulation validates the impact of prioritizing key tests in rare disease diagnosis: for acromegaly, IGF-1 and basal GH provide only screening clues, while the OGTT-GH suppression test is the gold standard for confirmation. It prompts clinicians that if initial tests (IGF-1/GH) are abnormal but nonspecific, OGTT must be mandated to avoid diagnostic delays. By contrasting probability changes before and after perturbation, it reinforces the clinical logic that "rare diseases require specific tests," optimizing diagnostic workflow efficiency. **Conclusion**: Abnormal results from the OGTT-GH suppression test are decisive evidence for acromegaly diagnosis. Their absence leads to diagnostic ambiguity, while supplementation significantly improves diagnostic specificity. Counterfactual simulation reveals the central role of key tests in rare disease differentiation, correcting clinicians' cognitive gaps.	Case 2: In the original data, clinicians may exclude Gitelman syndrome based solely on normal serum potassium and nonsignificant urine potassium, while AI can earlier identify the rare disease pattern of metabolic alkalosis mismatched with urine potassium excretion, unaffected by the "hypokalemia anchoring effect." After counterfactual perturbation, a bicarbonate level of 38 mmol/L and pH 7.6 clearly indicate metabolic alkalosis, prompting clinicians to reassess the possibility of renal tubular dysfunction. Value of Counterfactual Simulation: Overcoming Uncertainty: Clinicians initially believed "normal serum potassium rules out Gitelman syndrome," but counterfactual data drive re-evaluation of diagnostic logic. Correcting Cognitive Anchoring: The simulation targets clinicians' decision blind spots (e.g., undetected HCO3-), pushing them to shift from a "common disease" framework to complex pathophysiological analysis. Educational Significance: Reminds clinicians to screen for Gitelman syndrome (e.g., via genetic testing or magnesium/calcium urine tests) when acid-base imbalance exists despite normal serum potassium. Clinical Significance: Highlights the need for complete blood gas and electrolyte tests in hypokalemia/alkalosis patients to avoid misdiagnosis due to missed key indicators. **Conclusion**: This case demonstrates that counterfactual simulation, by supplementing critical evidence of metabolic alkalosis, helps clinicians overcome cognitive limitations of "hypokalemia" and "common disease priority," enabling more comprehensive evaluation of Gitelman syndrome. It underscores the tool's importance in rare disease differential diagnosis.	Case 3: The core feature of Gitelman syndrome is renal potassium wasting. Elevated urine potassium (>20 mmol/24h), when combined with hypokalemia (serum potassium 2.2 mmol/L), strongly supports renal tubular dysfunction (e.g., distal tubule NCC channel inactivation), distinguishing it from gastrointestinal potassium loss or inadequate intake. Metabolic alkalosis, a typical manifestation, is driven by renal hydrogen ion loss (via Na^+-H^+ exchange compensating for hypovolemia) and renin-angiotensin-aldosterone system (RAAS) activation related to hypomagnesemia. In the original diagnosis, missing urine potassium and HCO_3^- measurements may lead clinicians to default to "non-renal potassium loss" or "compensatory alkalosis," favoring common diseases (e.g., diarrhea, diuretic abuse). After perturbation, significantly elevated urine potassium and HCO_3^- directly indicate renal tubular dysfunction, consistent with Gitelman pathophysiology. Persistent hypokalemia (2.2 mmol/L) alongside elevated urine potassium creates a "hypokalemia-hyperuricosuria" paradox, a key marker distinguishing Gitelman syndrome from non-renal potassium loss. Value of Counterfactual Simulation: Cognitive Anchoring Correction: Perturbation targets the most uncertain indicators (undetected urine potassium and HCO_3^-), breaking the "common disease first" mindset through simulated abnormal values and forcing rare disease reassessment. Avoiding Premature Rare Disease Exclusion: Clinicians may underestimate Gitelman syndrome due to incomplete testing or assuming "nan values indicate normalcy." Counterfactual simulation, via data-driven abnormal values, highlights the diagnostic weight of key indicators, prompting early rare disease differential diagnosis. **Conclusion**: This case, by perturbing "nan" values of urine potassium and HCO_3^-, reveals the critical paradox of Gitelman syndrome (hypokalemia + hyperuricosuria + metabolic alkalosis + no hypertension), significantly increasing its diagnostic probability. The value lies in quantifying the potential impact of undetected indicators, assisting clinicians in overcoming cognitive limitations and systematically improving differential diagnosis processes.
Descrption	In the presented cases: 1. The true labels represent the actual disease status recorded in clinical practice. 2. For each indicator, the value before the arrow is the patient's actual test result (where "nan" indicates that the patient did not undergo that particular test), and the value after the arrow is the generated counterfactual indicator result. For missing indicators, the Δ change is calculated as the difference between the mean value of that indicator in the dataset and the counterfactual data. For other indicators, the Δ change is calculated as the difference between the original data and the counterfactual data. 3. The changes in prediction probabilities are obtained from a trained accurate AI model. The higher the score, the greater the likelihood. The cases respectively demonstrate the AI prediction results for the patient's original tests and the AI prediction results under counterfactual scenarios. Taking the first case below as an example: When the original tests were conducted, the model predicted the probability of a common disease to be 0.7770 and that of a rare disease to be 0.2230. After counterfactual perturbation, the probability of a common disease dropped sharply to 0.1297, while the probability of a rare disease increased to 0.8703, intuitively showing the reversing effect of supplementing key indicators on the diagnostic tendency. This case indicates that the reasonable supplementation of key missing indicators can significantly change the AI diagnostic tendency through counterfactual reasoning, providing a quantitative reference for clinicians to identify potential rare diseases.		

Figure 6. Prompting LLM and LLM response under three counterfactual scenarios