MedEinst: Benchmarking the Einstellung Effect in Medical LLMs through Counterfactual Differential Diagnosis

Wenting Chen[1], **Guolin Huang**[2], **Wenxuan Wang**[3], **Zhongrui Zhu**[4*]
[1]Department of Radiation Oncology, Stanford University,
[2]College of Computer Science and Software Engineering, Shenzhen University,
[3]Department of Computer Science, Renmin University of China,
[4]School of Computer Science and Technology, Xi'an Jiaotong University
wentchen@stanford.edu, zhongruizhu@stu.xjtu.edu.cn

Abstract

Despite achieving high accuracy on medical benchmarks, LLMs exhibit the Einstellung Effect in clinical diagnosis—relying on statistical shortcuts rather than patient-specific evidence, causing misdiagnosis in atypical cases. Existing benchmarks fail to detect this critical failure mode. We introduce **MedEinst**, a counterfactual benchmark with 5,383 paired clinical cases across 49 diseases. Each pair contains a control case and a "trap" case with altered discriminative evidence that flips the diagnosis. We measure susceptibility via Bias Trap Rate—probability of misdiagnosing traps despite correctly diagnosing controls. Extensive evaluation of 17 LLMs shows frontier models achieve high baseline accuracy but severe bias trap rates. Thus, we propose **ECR-Agent**, aligning LLM reasoning with Evidence-Based Medicine standards via two components: (1) Dynamic Causal Inference (DCI) performs structured reasoning through dual-pathway perception, dynamic causal graph reasoning across three levels (association, intervention, counterfactual), and evidence audit for final diagnosis; (2) Critic-Driven Graph & Memory Evolution (CGME) iteratively refines the system by storing validated reasoning paths in an exemplar base and consolidating disease-specific knowledge into evolving illness graphs. Data and code are publicly available at https://github.com/zhui711/MedEinst.

1 Introduction

Large Language Models (LLMs) (Achiam et al., 2023; Touvron et al., 2023) and LLM-based agents (Tang et al., 2024; Kim et al., 2024) achieve high performance on medical benchmarks (Jin et al., 2021). However, Kim et al. (2025) show these models exhibit the **Einstellung Effect**, relying on statistical shortcuts rather than logical reasoning. This causes models to prioritize common patterns over patient-specific evidence when encountering misleading features, ignoring key discriminative evidence. This effect is particularly problematic in differential diagnosis (DDx), where distinguishing between competing hypotheses depends on subtle symptomatic differences. Mitigating the Einstellung Effect in DDx is essential for deploying trustworthy clinical AI systems.

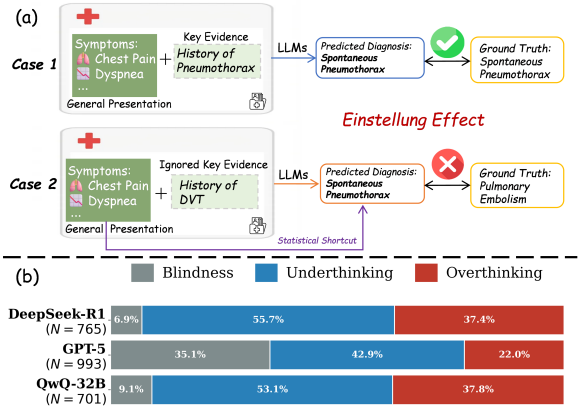

Figure 1: (a) Example of Einstellung Effect (b) Distribution of failure modes under the Einstellung Effect across reasoning LLMs, including **Blindness** (missing key evidence), **Underthinking** (insufficient reasoning), and **Overthinking** (rationalizing incorrect priors).

Although various medical benchmarks evaluate Med-LLMs (Singhal et al., 2023; Nori et al., 2023; Yan et al., 2024), they assess general medical capabilities rather than susceptibility to the Einstellung Effect. Existing benchmarks focus on knowledge evaluation (e.g., Medical QA on USMLE(Jin et al., 2021; Pal et al., 2022)) or clinical task performance (e.g., Clinical Summarization(Johnson et al., 2023) and Prognosis Prediction (Jiang et al., 2023; Chen et al., 2024)), testing static knowledge recall and standardized procedures. The Einstellung Effect manifests critically in DDx scenarios requiring identification of subtle discriminative features between similar diseases. Detecting this effect requires a **counterfactual evaluation design**:

*Corresponding Author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 39778–39798

presenting cases with similar symptoms but different diagnoses to assess whether models override pattern-based shortcuts for case-specific reasoning. However, current benchmarks lack such counterfactual scenarios. Thus, a specialized benchmark is needed to evaluate the Einstellung Effect in LLMs.

While current reasoning LLMs demonstrate strong logical capabilities, they remain susceptible to the Einstellung Effect in differential diagnosis. These models follow a "think-before-answer" paradigm but primarily establish simple symptom-disease associations rather than identifying discriminative evidence to disrupt pattern-based shortcuts. In Fig. 1, GPT-5 exhibits **blindness** in over 35% of error cases—completely ignoring key discriminative symptoms and defaulting to stereotypical diagnoses. Among cases where key symptoms are acknowledged, 43% involve **underthinking** (insufficient analysis) and 22% involve **overthinking** (motivated reasoning). These patterns reveal that current models lack structured mechanisms for rigorous evidence analysis. In contrast, real-world clinical practice follows Evidence-Based Medicine (EBM)(Sackett, 1997) framework: (1) **Problem Representation**—objectively reconstructing patient conditions; (2) **Acquire & Appraise**—actively seeking and verifying discriminative evidence; and (3) **Apply**—grounding diagnoses in verified evidence. Existing reasoning LLMs unfold reasoning linearly based on intuition, forcing a black-box "Symptoms → Diagnosis" mapping while neglecting the interpretable "**Symptoms → Evidence Verification → Diagnosis**" path. Therefore, constructing a reasoning framework grounded in EBM's cognitive architecture is imperative to mitigate the Einstellung Effect.

MedEinst: To bridge these gaps, we introduce **MedEinst**, a benchmark for evaluating the Einstellung Effect in medical LLMs via counterfactual differential diagnosis. MedEinst contains 5,383 paired clinical cases spanning 49 diseases across eight departments. To enable counterfactual evaluation, we employ a rigorous four-stage pipeline to generate the paired samples. Each pair consists of a *control* case and a minimally edited *trap* case: the trap case preserves most contextual evidence from the control case but replaces only the key discriminative evidence so that the correct diagnosis flips to a competing disease. This paired design creates counterfactual DDx scenarios in which superficial pattern matching strongly favors the original label, while correct diagnosis requires attending to the modified discriminative evidence. Using these pairs, we quantify susceptibility to the Einstellung Effect with **Bias Trap Rate**, the probability that a model—despite correctly solving the control case—misdiagnoses the trap case as the control label. We evaluate a broad set of 10 general and 5 medical-domain LLMs, as well as 2 LLM-based agents on MedEinst, and observe substantial Einstellung Effect errors across different models.

ECR-Agent: To mitigate the Einstellung Effect, we propose **ECR-Agent** (Evidence-based Causal Reasoning Agent), an agentic framework that emulates clinicians' EBM-grounded reasoning process through explicit discriminative evidence verification. ECR-Agent comprises two core components: (1) **Dynamic Causal Inference (DCI)** for structured diagnostic reasoning, and (2) **Critic-Driven Graph and Memory Evolution (CGME)** for accumulating clinical experience. The DCI module operationalizes the EBM framework through three stages. First, **dual-pathway perception** generates both intuitive differential diagnoses and an objective problem representation from patient symptoms, preventing premature diagnostic closure. Second, **dynamic causal graph reasoning** systematically seeks and verifies discriminative evidence through three progressive steps, each corresponding to a level in Pearl's causal hierarchy (Pearl and Mackenzie, 2018)—moving from observing patterns to actively testing hypotheses to counterfactual verification: (i) **Causal Graph Initialization** (*Association level*—observing correlations)—constructs a causal graph connecting observed symptoms, candidate diseases, and a pre-defined illness graph with prior illness knowledge to establish initial diagnostic hypotheses based on symptom-disease associations; (ii) **Forward Causal Reasoning** (*Intervention level*—testing what happens if we seek new evidence)—actively retrieves discriminative evidence from external knowledge bases as pivot nodes while incorporating typical supporting evidence as general nodes, then evaluates how each piece of evidence supports or refutes competing diagnoses to prevent underthinking; (iii) **Backward Causal Reasoning** (*Counterfactual level*—asking "what if this disease were true?")—performs counterfactual verification by identifying what evidence would be missing for each hypothesis, represented as shadow nodes that penalize incomplete diagnostic support and prevent overthinking. Third, the **evidence audit** module computes an evidence-based causal graph score for each candidate disease,

generates graph summary with disease-centric subgraphs, retrieves similar cases from an exemplar base, and produces the final diagnosis grounded in verified evidence rather than pattern matching. The CGME module enables experience accumulation across cases. Using a critic model, it iteratively refines diagnostic predictions until correctness is achieved, then stores: (1) case-level experience—the complete reasoning trace in the exemplar base for future retrieval; and (2) illness-level experience—merging and refining causal subgraphs across cases into consolidated illness graphs that capture refined discriminative patterns for each disease. Our contributions are as follows:

- We propose **MedEinst**, the first benchmark for evaluating the Einstellung Effect in medical LLMs, and introduce a novel metric revealing substantial model susceptibility.
- We propose **ECR-Agent**, an evidence-based framework to systematically verify discriminative evidence and accumulate clinical experience, mitigating the Einstellung Effect.
- Through extensive experiments, we demonstrate ECR-Agent's superiority and reveal current LLMs suffer from Einstellung Effect.

2 Related Work

2.1 Medical LLMs and Agents

LLMs have progressed from general medical assistants (Singhal et al., 2023; Achiam et al., 2023) passing USMLE exams to reasoning models using "think-before-answer" paradigms and LLM-based agents employing collaboration and retrieval. Agentic frameworks like MDAgents (Kim et al., 2024) and MedAgents (Tang et al., 2024) use multi-role debate, while RAG systems like MedGraphRAG (Wu et al., 2024) and PrimeKG (Chandak et al., 2023) incorporate Knowledge Graphs to reduce hallucinations. However, current models suffer from the Einstellung Effect (Alavi Naeini et al., 2023; Kim et al., 2025), using associative "Symptoms $\rightarrow$ Diagnosis" mappings instead of systematically verifying discriminative evidence. This leads models to favor statistical shortcuts over patient-specific evidence, with multi-agent collaboration potentially amplifying Consensus Bias (Schmidgall et al., 2024a). We therefore introduce ECR-Agent, an Evidence-Based Medicine (EBM) agentic framework (Sackett, 1997) that systematically verifies discriminative evidence through structured "Symptoms $\rightarrow$ Evidence Verification $\rightarrow$ Diagnosis" reasoning.

2.2 Medical Benchmarks for LLMs

Benchmarks for medical LLMs have shifted from static knowledge recall to dynamic reasoning. Early datasets like MedQA (Jin et al., 2021) and PubMedQA (Jin et al., 2019) assess factual knowledge, while DDXPlus (Fansi Tchango et al., 2022) and AgentClinic (Schmidgall et al., 2024c) evaluate diagnostic processes. However, existing benchmarks typically employ Independent and Identically Distributed (I.I.D.) samples or standard clinical presentations. They lack adversarial and counterfactual designs required to expose the Einstellung Effect. High performance on these datasets may reflect statistical fitting rather than robust reasoning. Thus, we propose MedEinst, a benchmark to evaluate the Einstellung Effect in medical LLMs via counterfactual differential diagnosis.

3 MedEinst Benchmark

Overview. We introduce **MedEinst**, a benchmark to evaluate the Einstellung Effect in medical LLMs through counterfactual differential diagnosis via a four-stage construction pipeline (Fig. 2). Moreover, we propose the **Bias Trap Rate** to quantify how often models solve a control case but fail a minimally edited trap case due to superficial reasoning.

3.1 Problem Formulation

We formalize medical diagnosis as a mapping $f : \mathcal{X} \rightarrow \mathcal{Y}$,where $\mathcal{X}$ denotes the patient narrative space and $\mathcal{Y}$ is the label space of 49 pathologies. We define a **Counterfactual Pair** ($\mathbf{x}^c$,$\mathbf{x}^t$) consisting of: (1) **Control Case** ($\mathbf{x}^c$), a typical presentation where statistical priors align with the ground truth (GT) y_{gt}; and (2) **Trap Case** ($\mathbf{x}^t$), an adversarial variant generated via minimal modification. Crucially, x^t remains statistically similar to y_{gt} but logically implies a bias label y_{bias} due to specific discriminative evidence.

Definition 1 (Einstellung Effect). A model f exhibits the Einstellung Effect if and only if:

$$f(\mathbf{x}^c) = y_{gt} \quad \wedge \quad f(\mathbf{x}^t) = y_{gt} \tag{1}$$

This implies that while the model demonstrates fundamental diagnostic competence (evidenced by success on the control case), it fails to rectify its

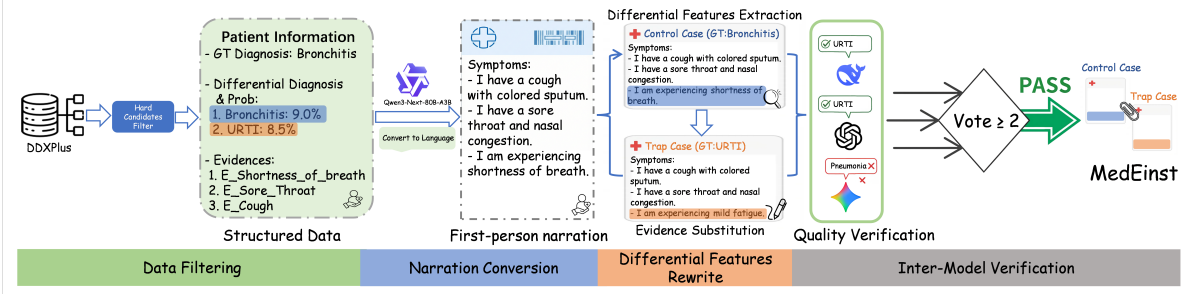

Figure 2: **Data construction of MedEinst** with four-stage process: (1) **Data Filtering** for hard candidates, (2) **Narration Conversion** to natural language, (3) **Differential Features Rewrite** for trap case generation, and (4) **Inter-Model Verification** for quality control.

prior intuition when confronted with the discriminative features in the trap case, rigidly persisting with the original diagnosis.

3.2 Benchmark Construction

Data Filtering. We collect 226,814 samples covering 49 pathologies from the DDXPlus dataset (Fansi Tchango et al., 2022) $\mathcal{D}src$ and filter for "Hard Candidates" where evidence-based reasoning is strictly necessary. Specifically, we select samples where the probability gap between the ground truth diagnosis y_{gt} and the top competing diagnosis y_{bias} is less than 0.5%. Relying on the original DDXPlus system, these probabilities represent the posterior probabilities calculated by their commercial diagnostic engine, based on epidemiological incidence rates (priors) and symptom likelihood ratios mapped from a massive medical knowledge base. This ensures that prior probabilities alone cannot distinguish between diagnoses and forcing the model to perform evidence-based differential diagnosis.

Narration Conversion. To simulate real-world clinical scenarios, we transform structured feature sets $\mathbf{s}$ into first-person natural language narratives $\mathbf{x}$ that capture the unstructured and noisy characteristics of actual medical records.

Differential Features Rewrite. This module precisely induces the Einstellung trap while maintaining clinical validity. To prevent hallucination, we ground our generation in the DDXPlus Knowledge Base ($\mathcal{K}$) rather than using standard rewriting. Specifically, we first perform *Differential Features Extraction* to identify the key discriminative features k_{gt} that distinguish y_{gt} from y_{bias}. Second, *Trap Information* generation (k_{trap}) strictly derives misleading evidence from the bias disease knowledge base K_{bias}. Finally, *Evidence Substitution* uses an LLM to replace k_{gt} with k_{trap}, generating x^t. This ensures the trap case logically points to y_{bias} while preserving all other contextual information from the control case.

Inter-Model Verification. To ensure high-quality pairs, we employ an "LLM-as-a-Judge" committee $\mathcal{J} = \{\text{GPT-5}, \text{DeepSeek-R1}, \text{Gemini-2.5-Pro}\}$ to assess each pair (x^c, x^t) across three dimensions: diagnostic correctness verifies whether x^t logically points to y_{bias}, medical plausibility assesses alignment with real-world medical logic, and narrative fluency evaluates text coherence (See Appendix B.2 for details). A pair is included in MedEinst $\mathcal{S}_{final}$ only if at least two judges vote positively on diagnostic correctness. As shown in Appendix Fig. 7, selected trap cases maintain high plausibility and fluency comparable to control cases, ensuring performance drops stem from reasoning failures rather than textual artifacts.

3.3 Dataset Statistics

MedEinst contains 5,383 counterfactual pairs of clinical narratives (10,766 cases total) covering 49 pathologies, derived from the DDXPlus test split to avoid data leakage. To provide an additional training set, we process and verify 10,689 pairs from the DDXPlus training split.

3.4 Quality Control

To ensure clinical validity in MedEinst, we implemented a rigorous quality control process involving four board-certified physicians with over 8 years of clinical experience. Our evaluation examined a stratified random sample of 1,500 counterfactual pairs (27.9% of the dataset). We developed a standardized scoring protocol evaluating seven binary quality dimensions: clinical plausibility of both control and trap cases, logical consistency of discriminative features, appropriateness of diagnoses, minimality of edits, and absence of artifactual pat-

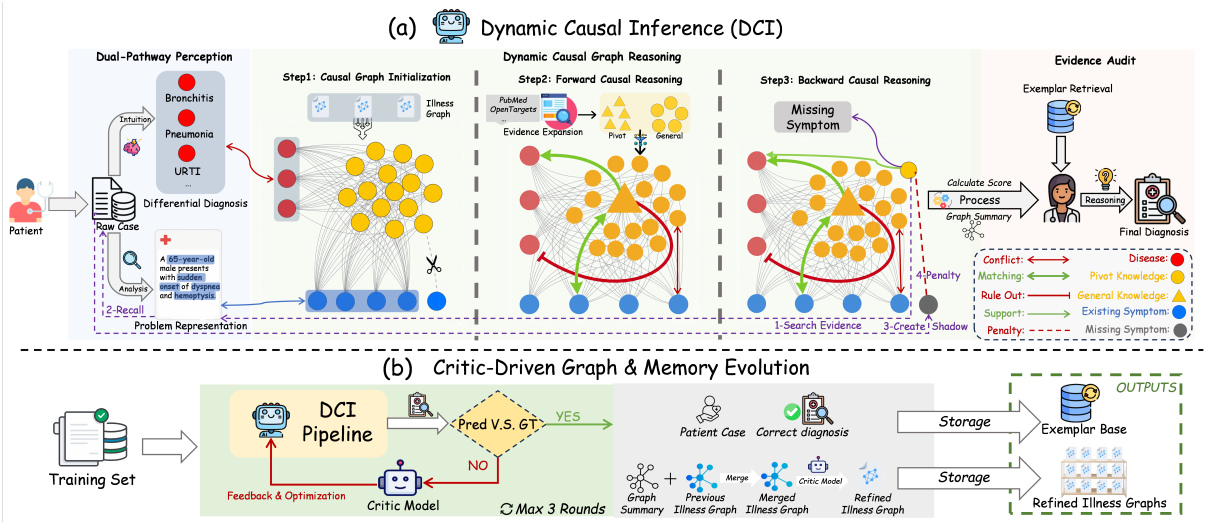

Figure 3: **ECR-Agent**, aligning LLM reasoning with Evidence-Based Medicine via two parts: (a) **Dynamic Causal Inference (DCI)** performs structured reasoning via *dual-pathway perception*, *dynamic causal graph reasoning* across three levels, and *evidence audit* for final diagnosis. (b) **Critic-Driven Graph & Memory Evolution (CGME)** iteratively refines the system by storing validated reasoning paths in an exemplar base and consolidating disease-specific knowledge into evolving illness graphs.

terns. Physicians evaluate each dimension through yes/no responses, with pairs satisfying all dimensions considered valid. The quality assessment yielded strong results, with 96.1% of evaluated pairs meeting our thresholds. Dimension-specific quality rates ranged from 94.3% to 98.2%. Inter-rater reliability analysis produced a Fleiss' kappa of 0.79, indicating substantial agreement. Pairs failing thresholds (3.9%) were either revised (2.1%) or excluded (1.8%) to maintain benchmark integrity.

3.5 Evaluation Metrics

To quantify the Einstellung Effect, we first prompt the model to generate diagnostic results for all counterfactual pairs (x^c, x^t). Then, we evaluate performance using three specific metrics based on the set of samples $S_{correct\_control}$ where the model correctly diagnosed the control case $(f(x^c) = y_{gt})$. **Baseline Accuracy** $(Acc_{base} = |S_{correct\_control}|/N_{total})$ establishes the model's fundamental diagnostic capability. **Robust Accuracy** $(Acc_{rob} = \sum_{i=1}^{N} \mathbb{I}(f(x_i^c) = y_{gt} \wedge f(x_i^t) = y_{bias})/N_{total})$ measures the proportion of pairs where the model correctly predicts both the control and trap cases. Finally, our primary metric, **Bias Trap Rate** $(R_{bias} = \sum_{i \in S_{correct\_control}} \mathbb{I}(f(x_i^t) = y_{gt})/|S_{correct\_control}|)$, calculates the conditional probability that a capable model falls into the trap given that the model possesses the fundamental diagnostic capability. N_{total} denotes the number of counterfactual pairs. A formal mathematical mapping of these outcomes—distinguishing Rigid Reversion from Stochastic Errors—along with a granular raw count breakdown, is provided in Appendix D.7.

4 ECR-Agent Framework

Overview. To mitigate the Einstellung Effect, we propose the ECR-Agent framework to align LLM reasoning with the rigorous verification standards of EBM (Fig. 3). ECR-Agent comprises two synergistic components: (1) Dynamic Causal Inference (DCI), which performs structured diagnostic reasoning through dual-pathway perception, a three-level causal graph verification process (spanning association, intervention, and counterfactual levels), and evidence audit; and (2) Critic-Driven Graph and Memory Evolution (CGME), which facilitates continuous improvement by refining diagnostic outputs and accumulating clinical experience into dynamic knowledge bases (Algorithm 2 and Appendix C).

4.1 Critic-Driven Graph & Memory Evolution

To accumulate diagnostic experience, we execute the DCI pipeline on the training set D_{train} and introduce a critic model M_{critic} (GPT-5) to orchestrate iterative refinement (Fig. 3 (b)). For each training case where the base model's prediction diverges from GT label, M_{critic} provides corrective feedback to optimize the reasoning path (maximum

3 rounds). Upon achieving correct diagnosis, the validated graph summary is merged with existing **illness graphs** $\mathcal{G} = \{G_y | y \in \mathcal{Y}\}$—a collection of disease-specific causal graphs initialized with the first graph summary—and further refined by the critic model to consolidate disease-level knowledge. Simultaneously, validated reasoning trajectories $(\mathbf{x}, y_{gt}, \text{Path})$ are stored in an **Exemplar Base** ($\mathcal{M}$) for case-based retrieval during inference.

4.2 The Dynamic Causal Inference (DCI)

4.2.1 Dual-Pathway Perception

To implement EBM's first principle of objective problem representation, we decouple statistical priors from factual observation through two parallel pathways. Firstly, the *intuitive pathway* generates Top-k candidate diagnoses $D_{set} = \{d_1, ..., d_k\}$ via Chain-of-Thought prompting, capturing pattern-based hypotheses. Secondly, the *analytic pathway* produces a *problem representation* that objectively summarizes key case features independent of diagnostic assumptions. From this representation, we extract structured patient observations $P_{obs} = \{p_1, ..., p_m\}$ and explicitly categorize each observation's status $s(p)$ as *Present* (affirmed), *Absent* (negated), or *Missing* (unmentioned). This dual-pathway design forces the model to acknowledge objective clinical facts before forming diagnostic conclusions, preventing premature closure driven by superficial pattern matching.

4.2.2 Dynamic Causal Graph Reasoning (DCGR)

DCGR aligns with Pearl's causal hierarchy through three levels: (1) **Causal Graph Initialization** (association) connects symptoms P_{obs} with candidates D_{set} via illness graphs $\mathcal{G}$; (2) **Forward Causal Reasoning** (intervention) retrieves and evaluates discriminative evidence; (3) **Backward Causal Reasoning** (counterfactual) penalizes hypotheses via expected-but-absent "shadow nodes".

Causal Graph Initialization. To establish initial diagnostic hypotheses based on observed correlations, we construct a causal graph integrating patient observations with disease knowledge. For each candidate $d \in D_{set}$, we retrieve its illness graph $G_{ill}^{(d)} = (V_d, V_p, V_k; E)$ from $\mathcal{G}$, where V_d, V_p, V_k represent disease, symptom, and knowledge nodes, and E denotes their relationships. We perform merge-or-prune based on embedding similarity between observations P_{obs} and V_p, retaining relevant nodes and merging novel observations, yielding the contextualized initial graph G_{ill}.

Forward Causal Reasoning. To prevent underthinking by actively seeking comprehensive discriminative evidence, we simulate the intervention: "What happens if we actively seek new evidence to differentiate among competing diseases?" We retrieve medical knowledge from external sources (PubMed, OpenTargets) and extract: (1) *pivot nodes* V_a—discriminative evidence differentiating diseases; (2) *general nodes* V_b—typical supporting evidence. We expand G_{ill} with these nodes: $G'_{ill} = G_{ill} \cup (V_a, V_b)$. Using Qwen3-32B, we identify 5 causal relations: *conflict*, *matching*, *rule out*, *support*, and *penalty*. For $V_p \leftrightarrow V_k$, we classify as conflict or matching; for $V_d \leftrightarrow V_k$, as rule out or support, producing the refined graph G'_{ill}.

Backward Causal Reasoning. To prevent overthinking and motivated reasoning, we perform counterfactual verification asking: "If disease d were true, what evidence should we observe?" For each d, we trace backward to identify supporting knowledge nodes $V_k^{(d)}$ and expected symptom nodes $V_p^{(d)}$. When knowledge node $v_k \in V_k^{(d)}$ lacks matching observations in P_{obs}, we trigger counterfactual verification (purple dashed line), re-examining the case text x. If evidence remains unverified, we instantiate a *shadow node* v_s (grey node) with a penalty edge to d, yielding a final causal graph $G^{\dagger}_{ill}$. Shadow nodes explicitly penalize hypotheses lacking expected evidence, ensuring diagnoses are grounded in verified evidence.

4.2.3 Evidence Audit

Graph Scoring and Summary: To quantify evidential support, we calculate an evidence-based causal graph score $S(d)$ for each candidate d: $S(d) = w_m N_{match}(d) - w_c N_{conf}(d) - w_s N_{shadow}(d)$, where $N_{match}(d)$, $N_{conf}(d)$, and $N_{shadow}(d)$ count edges with matching, conflict, and penalty relations, respectively, and w_m, w_c, w_s are weighting hyperparameters. We then generate a *Graph Summary* by reorganizing the causal graph $G^{\dagger}_{ill}$ into k disease-centric subgraphs, each centered on a candidate diagnosis. This reorganization preserves all graph information while structuring evidence around each hypothesis to facilitate evidence auditing.

ECR-Agent then integrates three information streams to derive the final diagnosis y^*: (1) *intuition*—initial reasoning from dual-pathway perception; (2) *evidence*—graph summary and scores $S(d)$; (3) *experience*—similar cases retrieved from

Table 1: Performance comparison of current LLMs and LLM-based Agents.

Model	Size	Baseline Acc (↑)	Robust Acc (↑)	Bias Trap Rate (↓)
Open-Source LLMs				
Kimi-k2	-	47.82	12.46	47.12
DeepSeek-R1	-	42.20	11.32	46.12
ZhipuAI/GLM-4.6	-	39.65	11.25	47.63
Qwen/Qwen3-14B	14B	44.12	11.28	54.19
Qwen/QwQ-32B	32B	41.05	11.14	44.88
Qwen/Qwen3-32B	32B	40.25	11.86	43.46
Qwen/Qwen3-235B-A22B	235B	40.51	11.40	43.32
Proprietary LLMs				
GPT-5	-	<u>54.30</u>	<u>15.78</u>	51.87
Gemini-2.5-pro	-	53.58	10.97	60.90
Claude-Sonnet-4.5	-	42.09	12.36	42.98
Medical-Specific LLMs				
Lingshu-7B	7B	14.93	3.20	<u>36.74</u>
Llama3-Med42-8B	8B	6.51	1.19	44.00
MedGemma-27B-text-it	27B	40.92	11.68	53.88
Baichuan-M2-32B	32B	45.03	7.18	66.10
Lingshu-32B	32B	27.68	6.17	54.20
LLM-based Agents				
MDAgent (Qwen3-32B)	32B	29.70	10.34	40.34
DyLAN (Qwen3-32B)	32B	32.11	8.11	41.69
ECR-Agent (Qwen3-32B)	**32B**	**69.49**	**24.21**	**33.75**

exemplar base $\mathcal{M}$. This holistic audit ensures the diagnosis is grounded in verified evidence rather than pattern-based biases.

5 Experiments

5.1 Evaluation Baselines

We compare ECR-Agent against three baseline types: 1) **General LLMs:** state-of-the-art proprietary (GPT-5, Claude-Sonnet-4.5, Gemini-2.5-Pro) and open-source LLMs (DeepSeek-R1, Qwen3-32B, QwQ-32B); 2) **Medical LLMs:** Lingshu-7B, Llama3-Med42-8B, MedGemma-27B-text-it, Baichuan-M2-32B and Lingshu-32B; 3) **LLM-based Agent:** MDAgent (Kim et al., 2024) and DyLAN (Liu et al., 2024b).

5.2 Overall Performance Comparison

Table 1 reveals a striking gap between diagnostic capability and robustness. While frontier models like GPT-5 and Gemini-2.5-Pro achieve the highest baseline accuracy (54.30% and 53.58%), they exhibit disproportionately high Bias Trap Rates (>50%), indicating a fundamental trade-off where models that better fit general medical distributions develop stronger priors that aggressively filter out low-probability counter-evidence (Perceptual Blindness, Fig. 1), making them more susceptible to Einstellung traps than weaker models. Further analyses of error directionality and prediction confidence confirm that these failures reflect true cognitive rigidity rather than random calibration errors or uncertainty (see Appendix D.6). Agent frameworks like MDAgent (multi-role debate) and DyLAN (dynamic agent selection) show low robust accuracy (~8-10%) and high trap rates due to noise amplification, where dynamic interaction topologies merely reinforce the dominant statistical prior (Consensus Bias) rather than correcting it—DyLAN's strategy of selecting "high-contribution" agents exacerbates this by favoring agents that align with the incorrect group consensus. In contrast, ECR-Agent achieves substantial improvements (69.49% baseline accuracy, 24.21% robust accuracy, 33.75% bias trap rate). Statistical validation via bootstrap resampling ($K = 1,000$) confirms these gains: the 95% Confidence Interval (CI) for ECR-Agent's robust accuracy is [23.07%, 25.35%], which completely exceeds GPT-5's CI of [14.80%, 16.76%]. A paired bootstrap test demonstrates that ECR-Agent significantly outperforms GPT-5 ($p < 0.001$), empirically validating that resolving the Einstellung Effect requires a paradigm shift from statistical fitting (probability) to causal verification (evidence).

To ensure that ECR-Agent's gains are not confined to counterfactual scenarios, we further eval-

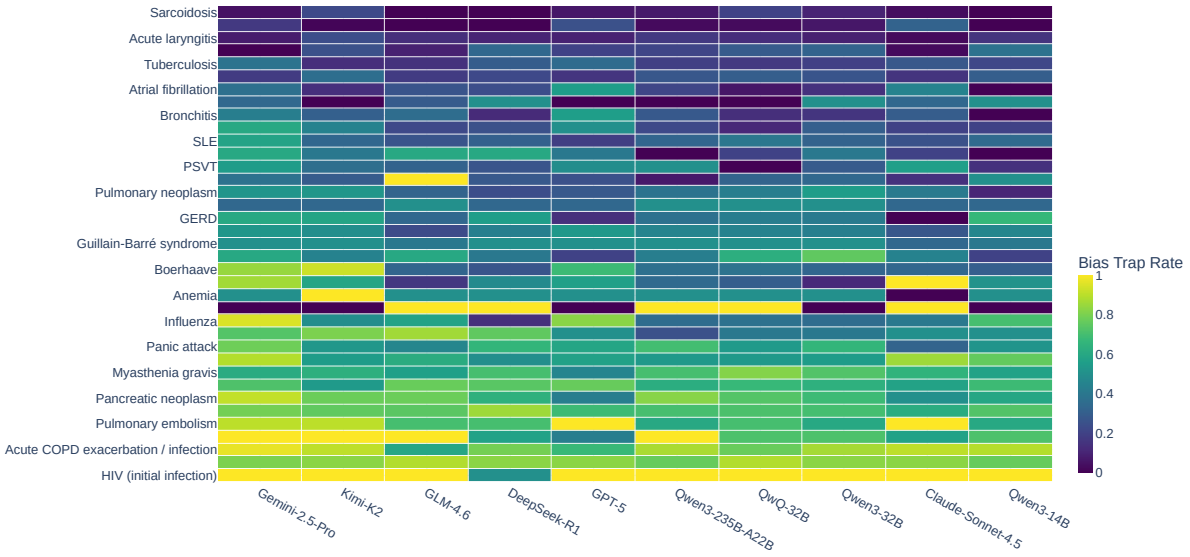

Figure 4: Bias Trap Rate heatmap across diseases. The clustering indicates that models learn spurious correlations for common diseases (e.g., Pneumonia), leading to consistent bias.

uate its generalizability on standard clinical presentations using the established MedChain (Liu et al., 2024a) benchmark. As shown in Table A1 (detailed in Appendix D.4), ECR-Agent achieves a competitive diagnostic accuracy of 46.09%, significantly outperforming general agentic frameworks such as DyLAN (38.63%) and MDAgent (39.73%). Remarkably, ECR-Agent performs on par with the domain-specific MedChain-Agent (48.07%) without any dataset-specific tuning. These results underscore that transitioning from probabilistic pattern matching to structured evidence verification enhances reasoning robustness without compromising—and indeed bolstering—general clinical efficacy across diverse real-world domains.

5.3 Ablation Study

We conduct an ablation study on ECR-Agent (Qwen3-32B as the base model) to evaluate the module effectiveness. In Table 2, adding DCI substantially improves Base Accuracy from 40.25% to 55.49%, showing the effectiveness of structured causal reasoning. Further incorporating CGME yields additional significant gains to 69.49% Base Accuracy and reduces Trap Rate to 33.75%, proving the critical role of experience accumulation.

5.4 Disease-Specific Analysis

Fig. 4 reveals heterogeneity in Bias Trap Rates across diseases, exposing the structural nature of

Table 2: **Ablation Study on ECR-Agent components.**

DCI	CGME	Base Acc (↑)	Rob Acc (↑)	Trap Rate (↓)
		40.25	11.86	43.46
✓		55.49	19.94	38.32
✓	✓	**69.49**	**24.21**	**33.75**

the Einstellung Effect: a systematic "High-Bias Cluster" emerges in diseases like *Pulmonary Embolism* and *Initial HIV Infection* (bottom rows) whose presentations overlap with high-prevalence distractors (e.g., Flu, Anxiety), where LLMs learn spurious correlations between generic symptoms and statistically probable diagnoses while ignoring key discriminative evidence. This failure persists across all architectures, e.g. reasoning-optimized (DeepSeek-R1) and massive-scale LLMs (Qwen3-235B), showing CoT capabilities as pattern matchers that collapse when diagnosis requires overriding priors with specific evidence.

5.5 Scaling Laws vs. Einstellung Effect

Fig. 5 reveals the limitations of scaling in robust medical reasoning. Rather than observing an emergence of robustness, we observe a phenomenon akin to Inverse Scaling (McKenzie et al., 2023): frontier models like GPT-5 and Gemini-2.5-Pro occupying a "High Capability, High Bias" region where scaling improves baseline diagnostic capability (ACC_{base}) but paradoxically exacerbates Ein-

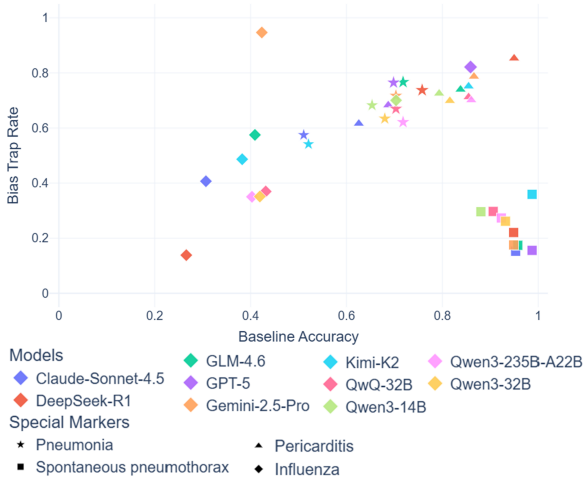

Figure 5: Baseline Accuracy vs. Bias Trap Rate.

stellung susceptibility. We term this phenomenon **"Stronger Priors, Stronger Blindness"**—an empirical observation within our 49-pathology dataset where larger models capture statistical regularities so effectively they become overconfident in initial intuitions, making it harder to override diagnoses when presented with subtle counter-evidence—a trend evident across model tiers (e.g., Gemini-2.5-Pro achieves superior baseline accuracy yet exhibits a 60.90% bias trap rate, significantly higher than less capable models). These findings demonstrate the Einstellung Effect as a fundamental cognitive failure mode that persists with scale, necessitating architectural interventions like ECR-Agent that decouple evidence verification from probabilistic generation.

To investigate whether the observed Einstellung bias is merely a byproduct of prompt sensitivity, we conduct a robustness stress test by augmenting the Zero-shot CoT prompt with an explicit warning against misleading distractors. Paradoxically, as shown in Table A2 (and detailed in Appendix D.5), the explicit warning exacerbates the bias: while it marginally improves baseline accuracy, the Bias Trap Rate significantly surges (e.g., from 43.46% to 60.50% for Qwen3-32B). This failure to override the bias through instruction reinforces our *"Stronger Priors, Stronger Blindness"* finding—the warning inadvertently strengthens the model's reliance on statistical priors, which then more aggressively suppress contradictory evidence. Such resistance confirms that the Einstellung Effect is rooted in an **intrinsic cognitive rigidity** of probabilistic pattern matchers, which cannot be mitigated by superficial prompting and necessitates the structural intervention of ECR-Agent.

6 Conclusion

We introduced MedEinst, the first counterfactual benchmark exposing the Einstellung Effect in medical LLMs, revealing that frontier models achieve high baseline accuracy yet remain severely susceptible to statistical shortcuts. We proposed ECR-Agent, which aligns LLM reasoning with Evidence-Based Medicine through structured causal inference and knowledge evolution.

Limitations

While MedEinst includes 5,383 counterfactual pairs, it currently covers only 49 common pathologies across eight departments. Although these diseases represent high-frequency diagnostic scenarios in emergency medicine, they constitute a small fraction of the vast medical ontology (e.g., ICD-10). Consequently, the manifestation of the Einstellung Effect in rare diseases or complex comorbidities remains to be fully explored. We view MedEinst as a foundational proof-of-concept, paving the way for future benchmarks to expand into broader disease taxonomies.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Saeid Alavi Naeini, Raeid Saqur, Mozhgan Saeidi, John Giorgi, and Babak Taati. 2023. Large language models are fixated by red herrings: Exploring creative problem solving and einstellung effect using the only connect wall dataset. *Advances in Neural Information Processing Systems*, 36:5631–5652.

Payal Chandak, Kexin Huang, and Marinka Zitnik. 2023. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10(1):67.

Canyu Chen, Jian Yu, Shan Chen, Che Liu, Zhongwei Wan, Danielle Bitterman, Fei Wang, and Kai Shu. 2024. Clinicalbench: Can llms beat traditional ml models in clinical prediction? *arXiv preprint arXiv:2411.06469*.

Arsene Fansi Tchango, Rishab Goel, Zhi Wen, Julien Martel, and Joumana Ghosn. 2022. Ddxplus: A new dataset for automatic medical diagnosis. *Advances in neural information processing systems*, 35:31306–31318.

Lavender Yao Jiang, Xujin Chris Liu, Nima Pour Nejatian, Mustafa Nasir-Moin, Duo Wang, Anas Abidin, Kevin Eaton, Howard Antony Riina, Ilya Laufer, Paawan Punjabi, and 1 others. 2023. Health system-scale language models are all-purpose prediction engines. *Nature*, 619(7969):357–362.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 2567–2577.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, and 1 others. 2023. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.

Jonathan Kim, Anna Podlasek, Kie Shidara, Feng Liu, Ahmed Alaa, and Danilo Bernardo. 2025. Limitations of large language models in clinical problem-solving arising from inflexible reasoning. *Scientific reports*, 15(1):39426.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik S Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae W Park. 2024. Mdagents: An adaptive collaboration of llms for medical decision-making. *Advances in Neural Information Processing Systems*, 37:79410–79452.

Jie Liu, Wenxuan Wang, Zizhan Ma, Guolin Huang, Yihang Su, Kao-Jung Chang, Wenting Chen, Haoliang Li, Linlin Shen, and Michael Lyu. 2024a. Medchain: Bridging the gap between llm agents and clinical practice with interactive sequence. *arXiv preprint arXiv:2412.01605*.

Zijun Liu, Yanzhe Zhang, Peng Li, Yang Liu, and Diyi Yang. 2024b. A dynamic llm-powered agent network for task-oriented agent collaboration. In *First Conference on Language Modeling*.

Ian R McKenzie, Alexander Lyzhov, Michael Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Aaron Kirtland, Alexis Ross, Alisa Liu, and 1 others. 2023. Inverse scaling: When bigger isn't better. *arXiv preprint arXiv:2306.09479*.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Judea Pearl and Dana Mackenzie. 2018. *The book of why: the new science of cause and effect*. Basic books.

Jonathan G Richens, Ciarán M Lee, and Saurabh Johri. 2020. Improving the accuracy of medical diagnosis with causal machine learning. *Nature communications*, 11(1):3923.

David L Sackett. 1997. Evidence-based medicine. *Seminars in perinatology*, 21(1):3–5.

Samuel Schmidgall, Carl Harris, Ime Essien, Daniel Olshvang, Tawsifur Rahman, Ji Woong Kim, Rojin Ziaei, Jason Eshraghian, Peter Abadir, and Rama Chellappa. 2024a. Addressing cognitive bias in medical language models. *arXiv preprint arXiv:2402.08113*.

Samuel Schmidgall, Carl Harris, Ime Essien, Daniel Olshvang, Tawsifur Rahman, Ji Woong Kim, Rojin Ziaei, Jason Eshraghian, Peter Abadir, and Rama Chellappa. 2024b. Evaluation and mitigation of cognitive biases in medical language models. *npj Digital Medicine*, 7(1):295.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. 2024c. Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments. *arXiv preprint arXiv:2405.07960*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2024. Medagents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 599–621.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Junde Wu, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min Xu, Filippo Menolascina, and Vicente Grau. 2024. Medical graph rag: Towards safe medical large language model via graph retrieval-augmented generation. *arXiv preprint arXiv:2408.04187*.

Lawrence KQ Yan, Qian Niu, Ming Li, Yichao Zhang, Caitlyn Heqi Yin, Cheng Fei, Benji Peng, Ziqian Bi, Pohsun Feng, Keyu Chen, and 1 others. 2024. Large language model benchmarks in medical tasks. *arXiv preprint arXiv:2410.21348*.

Appendix

Abstract. This appendix provides supplementary materials for the MedEinst benchmark and the ECR-Agent framework.

Appendix A details the methodological algorithms for benchmark construction and agent inference, along with the causal graph schema and evaluation metrics.

Appendix B provides a comprehensive analysis of the MedEinst benchmark, including clinical specialty distribution, quality assurance protocols, and dataset statistics.

Appendix C outlines the implementation details, including experimental settings and baseline configurations.

Appendix D presents additional empirical analyses, focusing on detailed failure modes and the capability-robustness gap.

Appendix E offers a concrete case study (Case 100473) to qualitatively demonstrate the reasoning trace and interpretability of our approach.

Appendix F extends the discussion on theoretical grounding, mapping our framework to the Causal Hierarchy and contrasting it with existing paradigms.

Appendix G outlines future work, addressing complex clinical realities such as rare diseases, multimorbidity, and longitudinal scenarios.

Appendix H displays raw data samples illustrating the input format.

Appendix I lists the detailed prompts used for data construction and the agent reasoning pipeline.

A Methodological Details

A.1 MedEinst Construction Algorithm

Algorithm 1 outlines the rigorous four-stage pipeline employed to construct the MedEinst benchmark. The process begins with Data Filtering to select "Hard Candidates" where statistical shortcuts fail. It then proceeds to Narration Conversion and Differential Features Rewrite, transforming structured data into natural language and injecting adversarial traps based on knowledge base from DDXPlus (Fansi Tchango et al., 2022). Finally, Inter-Model Verification serves as a quality control filter, ensuring that the generated trap cases are medically plausible.

Algorithm 1 Construction Pipeline of **MedEinst**

Input: Source dataset $\mathcal{D}_{src}$, Knowledge Base $\mathcal{K}$ (DDXPlus), LLM Judge Committee $\mathcal{J}$; Threshold $\epsilon = 0.5\%$.
Output: Paired Counterfactual Benchmark $\mathcal{S}_{final}$.

Initialize $\mathcal{S}_{final} \leftarrow \emptyset$
for each sample $(\mathbf{s}, y_{gt}, P) \in \mathcal{D}_{src}$ **do**
  Step 1: Data Filtering
  if $|P(y_{gt}) - P(y_{bias})| < \epsilon$ **then**
    Step 2: Narration Conversion
    $x^c \leftarrow \text{LLM}(\mathbf{s})$
    Step 3: Differential Features Rewrite
    Retrieve $K_{gt}, K_{bias} \leftarrow \text{Query}(\mathcal{K}, \{y_{gt}, y_{bias}\})$
    $k_{gt} \leftarrow \text{LLM}(x^c, K_{gt}, K_{bias})$
    $k_{trap} \leftarrow \text{LLM}(K_{bias}, k_{gt})$
    $x^t \leftarrow \text{LLM}(x^c, k_{trap}, k_{gt})$
    Step 4: Inter-Model Verification
    $V_{score} \leftarrow \sum_{j \in \mathcal{J}} \mathbb{I}(\text{LLM}_j(x^t, y_{bias}) = \text{Correct})$
    if $V_{score} \geq 2$ **then**
      $\mathcal{S}_{final} \leftarrow \mathcal{S}_{final} \cup \{(x^c, x^t, y_{gt}, y_{bias})\}$
    end if
  end if
end for
return $\mathcal{S}_{final}$

A.2 ECR-Agent Inference Algorithm

Algorithm 2 formally describes the complete workflow of the ECR-Agent, integrating both the training and inference phases. The algorithm first details the Critic-Driven Graph & Memory Evolution (CGME), where the system iteratively refines illness graphs and accumulates an exemplar base using critic feedback on the training set. Subsequently, it presents the Dynamic Causal Inference (DCI) pipeline used during inference, which orchestrates Dual-Pathway Perception, Dynamic Causal Graph Reasoning (across initialization, forward, and backward steps), and the final Evidence Audit to derive robust diagnoses for unseen cases.

A.3 Causal Reasoning Graph

Graph Schema Definition:

- **Patient Nodes**(V_P): Encode structured clinical observations extracted from problem representation. Crucially, we distinguish node status $s(p)$ into three states: *Present* (affirmed), *Absent* (negated), *Missing* (unmentioned).
- **Knowledge Nodes**(V_K): Encode disease-specific clinical entities (e.g., symptoms, biomarkers) distilled from literature. They are categorized into **General** (typical features) and **Pivot** (discriminators).

Merge-or-Prune operation

$$\text{Action}(p_{script}) = \begin{cases} \text{Merge}, & \text{if } \cos(\mathbf{e}_{p_{script}}, \mathbf{e}_{p_{obs}}) > \tau \\ \text{Prune}, & \text{otherwise} \end{cases} \tag{2}$$

where $\tau = 0.9$. This ensures G_{curr} only contains the patient's actual data while inheriting relevant causal structures from the illness graphs.

Algorithm 2 ECR-Agent Evolution & Inference Pipeline

Input: Training Set $\mathcal{D}_{train}$; New Case $\mathbf{x}_{new}$
Output: Refined Illness Graphs $\mathcal{G}_{refined}$; Exemplar Base $\mathcal{M}$; Diagnosis $d^\star$; Causal Reasoning Graph $G_{ill}^{(d^\star)}$

/* Critic-Driven Graph & Memory Evolution */
for sample $(\mathbf{x}, y_{gt}) \in \mathcal{D}_{train}$ **do**
  $t \leftarrow 0$
  while $t < 3$ **do**
    $(d_{pred}, G_{summary}) \leftarrow$ DCI_Pipeline$(\mathbf{x})$
    if $d_{pred} == y_{gt}$ **then**
      Load Previous Graph G_{prev} for y_{gt}
      $G_{merged} \leftarrow$ Merge$(G_{prev}, G_{summary})$
      break
    else
      ApplyCriticFeedback$(\mathbf{x})$
      $t \leftarrow t + 1$
    end if
  end while
  ▷ If loop ends without success, sample is discarded.
end for

/* Dynamic Causal Inference (DCI) Pipeline */
function DCI_PIPELINE$(\mathbf{x})$
  Dual-Pathway Perception
  $D_{set} \leftarrow$ IntuitivePathway(x)
  $P_{obs} \leftarrow$ AnalyticPathway(x)
  Dynamic Causal Graph Reasoning
  for candidate $d \in D_{top}$ **do**
    Load $G_{ill}^{(d)} = (V_p, V_k, E)$
    Step 1: Causal Graph Initialization
    $V_{init} \leftarrow \{p \in V_p \mid \text{Sim}(p, P_{obs}) > \tau\}$
    $G_{ill} \leftarrow$ Initialize(V_{init}, E)
    Step 2: Forward Causal Reasoning
    $V_k \leftarrow$ LiveSearch(d)
    ▷ Expand Pivot/General Nodes
    $G_{ill}^{'} \leftarrow G_{ill} \cup$ Link$(V_d, V_k) \cup$ Link(V_k, V_p)
    Step 3: Backward Causal Reasoning
    $\Delta_{miss} \leftarrow \{k \in V_k^{(d)} \mid k \notin P_{obs} \wedge$ IsExpected$(k)\}$
    $N_{shadow}^{(d)} \leftarrow \emptyset$
    for $k \in \Delta_{miss}$ **do**
      if ReExamine$(\mathbf{x}, k) ==$ Found **then**
        $P_{obs} \leftarrow P_{obs} \cup \{k\}$; Update $G_{ill}^{\dagger}$
      else
        $N_{shadow}^{(d)} \leftarrow N_{shadow}^{(d)} \cup \{k\}$
        ▷ Create Shadow Node
      end if
    end for
    ▷ This $G_{ill}^{\dagger}$ serves as the "Graph Summary"
    $Score(d) \leftarrow$ CalculateScore$(G_{ill}^{\dagger}, N_{shadow}^{(d)})$
  end for
  Evidence Audit
  $\mathcal{M}_{sim} \leftarrow$ RetrieveExemplars$(\mathcal{M}, P_{obs})$
  $d^\star \leftarrow$ LLM_Judge$(D_{top}, \{Score(d)\}, \mathcal{M}_{sim})$
  return $(d^\star, G_{ill}^{(d^\star)})$
end function

A.4 Evaluation Metrics

To precisely quantify the Einstellung Effect, we classify the model's predictions on paired samples (x^c, x^t) into three categories based on the intersection of their outcomes. Let $S_{correct\_control}$ denote the set of samples where the model correctly diagnoses the Control Case $(f(x^c) = y_{gt})$. We define the following metrics:

- **Baseline Accuracy (Acc_{base}):** Measures the fundamental diagnostic capability on standard clinical presentations.
$$Acc_{base} = \frac{|S_{correct\_control}|}{N_{total}} \tag{3}$$
- **Robust Accuracy (Acc_{rob}):** Measures the proportion of pairs where the model maintains correctness across both control and trap cases (Robust Success).
$$Acc_{rob} = \frac{\sum_{i=1}^{N} \mathbb{I}(f(x_i^c) = y_{gt} \wedge f(x_i^t) = y_{bias})}{N_{total}} \tag{4}$$
- **Bias Trap Rate (R_{bias}):** The core metric for the Einstellung Effect. It measures the conditional probability of fall in the trap given that the model possesses the fundamental diagnostic capability.
$$R_{bias} = \frac{\sum_{i \in S_{correct\_control}} \mathbb{I}(f(x_i^t) = y_{gt})}{|S_{correct\_control}|} \tag{5}$$

B MedEinst Benchmark Details

B.1 Clinical Specialty Analysis

To assess the clinical breadth and diversity of the **MedEinst Benchmark**, we categorized the 49 target pathologies into 10 distinct clinical specialties. Unlike rigid anatomical classifications (e.g., ICD-10), we adopted a clinical taxonomy based on medical specialties and triage departments. This approach better reflects real-world diagnostic workflows where pathologies presenting with overlapping symptoms are managed by specific domains.

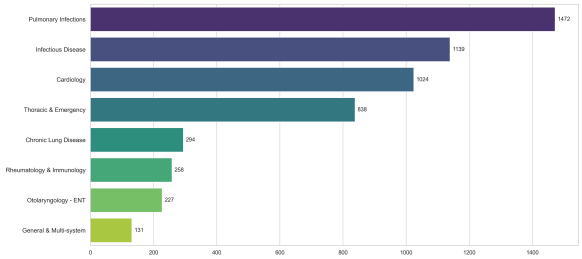

Figure 6: **Distribution of MedEinst Benchmark Pairs by Clinical Specialty.** The 5,383 test pairs are grouped into 10 categories based on standard clinical taxonomy. The high representation of Pulmonary and Cardiology cases reflects the dataset's focus on acute care scenarios where differential diagnosis is most critical.

B.2 Quality Assurance

To verify that our **Differential Features Rewrite** (Section 3.2) does not degrade the linguistic or clinical quality of the patient narratives, we analyzed the distribution of *Medical Plausibility* and *Narrative Fluency* scores assigned by the judge committee $\mathcal{J} = \{\text{GPT-5}, \text{DeepSeek-R1}, \text{Gemini-2.5-Pro}\}$.

Figure 7 presents the comparative analysis between **GOOD Cases** (successfully generated traps that passed verification) and **BAD Cases** (rejected traps).

- **Medical Plausibility**: The GOOD cases (green boxplots) maintain a high median score ($\approx 8.0/10$), statistically indistinguishable from the original clinical notes. This confirms that the injected *trap_info* aligns logically with the patient's context (e.g., age, gender, symptoms, antecedents).

- **Narrative Fluency**: The rewriting process preserves the natural flow of the text, with GOOD cases achieving a median fluency score of $\approx 8.3/10$. In contrast, BAD cases often exhibit disjointed insertions or grammatical inconsistencies, justifying their exclusion.

This quality audit confirms that the **Einstellung Effect** observed in our benchmark stems from the model's inability to process conflicting evidence, rather than poor data quality.

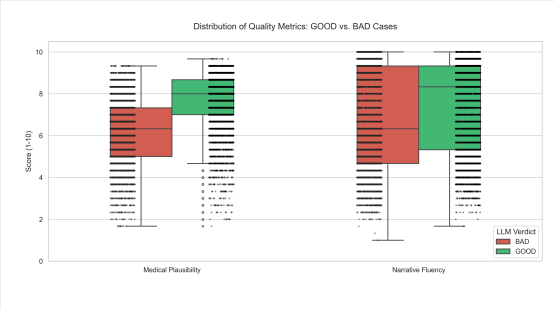

Figure 7: Distribution of quality metrics (Medical Plausibility and Narrative Fluency) for accepted (GOOD) versus rejected (BAD) trap cases. The high scores of accepted cases validate the effectiveness of our isomorphic rewriting protocol.

B.3 Dataset Statistics

We constructed **MedEinst** based on the DDXPlus dataset, strictly adhering to its original chronological split to prevent data leakage. The benchmark comprises two subsets:

- **Test Set (The Benchmark):** Derived from the DDXPlus test split, this set contains 5,383 counterfactual pairs of clinical narratives (totaling 10,766 cases) covering 49 pathologies. A unique feature of MedEinst is its *Paired Counterfactual* design: each Control Case (x^c) is paired with a Trap Case (x^t) that differs only in **Key Discriminative Features**, yet leads to a contradictory diagnosis (y_{gt} vs. y_{bias}). This design strictly decouples a model's statistical intuition from its logical reasoning capability.

- **Reference Set (Training Resource):** Derived from the DDXPlus training split, we processed and verified 10,689 pairs. This large-scale set is provided to support various research paradigms, including fine-tuning, few-shot learning, or RAG-based retrieval.

Selection Criteria. A sample pair is included in the final MedEinst benchmark $\mathcal{S}_{final}$ if and only if it receives a positive vote on Diagnostic Correctness from at least two judges. As shown in Appendix B.2, the selected trap cases maintain high medical plausibility and narrative fluency comparable to control cases. This rigorous verification ensures that performance drops in MedEinst stem from reasoning failures (Einstellung Effect) rather than textual artifacts or data noise.

C Implementation Details

C.1 Agent Configuration

To simulate a realistic clinical diagnosis scenario where physicians encounter unseen cases, all baseline models and agent frameworks operate under a Zero-shot Chain-of-Thought (CoT) setting. For our **ECR-Agent**, we maintain the same zero-shot input for fair comparison. Specifically, in the *Dual-Pathway Perception* phase, we configure the agent to generate the Top-k candidate diagnoses with $\boldsymbol{k = 5}$. This threshold was empirically selected to ensure sufficient coverage of potential differentials (including the ground truth and trap) while maintaining computational efficiency for the subsequent causal graph construction. Evidence Expansion is supported by structured queries to OpenTargets and PubMed APIs, functioning as an extension of the agent's analytic system.

C.2 Dataset and Sampling Strategy

We evaluate all methods on the MedEinst benchmark (5,383 pairs). To drive the Critic-Driven Graph & Memory Evolution, we utilized the MedEinst-Support set. To demonstrate the data efficiency of our framework, we did not employ the full support set. Instead, we curated a compact Balanced Seed Subset consisting of only 853 cases (approximately 8% of the available training data).This subset was constructed using a Capped Sampling Strategy: we randomly sampled a maximum of $\boldsymbol{N = 20}$ cases per pathology, while retaining all available samples for rare diseases. This lightweight selection ensures that the agent can initialize robust *Illness Graphs* and the *Exemplar Base* with minimal data consumption, highlighting the framework's capability to generalize from sparse but balanced clinical examples.

C.3 Computational Cost and Inference Latency

Deployed via the Qwen3-32B Cloud API, the ECR-Agent pipeline requires exactly 6 LLM calls per clinical case, resulting in a mean total inference latency of 122.28 seconds. This represents an computational overhead of approximately 66.60 seconds compared to the standard zero-shot CoT inference of the base Qwen3-32B model (55.68 seconds). However, this additional temporal cost yields substantial empirical gains: an absolute improvement of over 20% in Baseline Accuracy ($\boldsymbol{Acc_{base}}$) and an approximate 10% absolute reduction in the Bias Trap Rate ($\boldsymbol{R_{bias}}$). In the context of high-stakes differential diagnosis—where relying on statistical shortcuts for overlapping symptoms can mask fatal conditions (e.g., the Pulmonary Embolism vs. Spontaneous Pneumothorax scenario detailed in our Case Study)—we argue that an additional $\sim$1.1 minutes dedicated to systematic causal verification is a strictly necessary and clinically acceptable trade-off to ensure patient safety.

D Additional Experimental Analysis

To investigate the microscopic mechanisms and macroscopic characteristics of the Einstellung Effect, we conduct a multi-dimensional empirical analysis.

D.1 Detailed Failure Mode Analysis

To understand the cognitive failures behind Einstellung Traps, we conducted a fine-grained failure analysis on three representative models (DeepSeek-R1, GPT-5, QwQ-32B). We classify reasoning failures into three modes based on the model's interaction with the Key Discriminative Evidence. The classification was performed by a GPT-5 Auditor and **verified by human experts on a subset of data** (Cohen's $\boldsymbol{\kappa > 0.8}$).

As shown in Figure 1, the distribution reveals distinct cognitive deficits:

- **Blindness** Models completely fail to mention the key evidence in their CoT. This suggests that strong statistical priors filter out "unexpected" symptoms during the initial perception stage. *Our Solution (Dual-Pathway Perception):* We introduce **Dual-Track Perception**, forcing the explicit extraction of a structured Problem Representation to ensure all evidence is "seen".

- **Underthinking** Even when evidence is seen, models often default to the most likely candidate without rigorous falsification. *Our Solution (Causal Graph Reasoning):* We implement **Causal Graph Reasoning**. By constructing a patient-specific graph with Pivot Nodes, we structurally force bidirectional reasoning (Forward Support & Backward Exclusion) to prevent the dismissal of contradictory evidence.

- **Overthinking** Advanced models (e.g., GPT-5) engage in **Motivated Reasoning**, hallucinating mechanisms to force-fit contradictions into the incorrect diagnosis. *Our Solution (Evidence Audit):* We deploy an **Evidence Audit**. By performing Counterfactual Checks, the agent detects and penalizes such non-causal rationalizations, breaking the self-confirming loop.

D.2 Overall Performance Comparison

Table 1 presents the performance of various models and agent frameworks on MedEinst. We observe three critical phenomena:

1. The Capability-Robustness Gap. While frontier models like GPT-5 and Gemini-2.5-Pro demonstrate superior fundamental diagnostic capabilities ($\boldsymbol{Acc_{base}}$ of 54.30% and 53.58% respectively), their robustness remains disproportionately low, with $\boldsymbol{Acc_{rob}}$ hovering around 10%–15%. Alarmingly, these stronger models often exhibit higher susceptibility to Einstellung traps ($\boldsymbol{R_{bias}}$ 51%–61%). For instance, **Gemini-2.5-Pro**, despite its high capability, shows a significantly higher bias rate (60.90%) compared to **Claude-Sonnet-4.5** (42.98%). This implies that in adversarial contexts, high capability can paradoxically increase vulnerability to bias. This result reveals a counter-intuitive conclusion: **the very success of scaling in optimizing 'statistical fitting' paradoxically compromises 'differential diagnostic capability in dynamic contexts'. Rather than a failure of model capability, this represents a structural vulnerability where scaling exacerbates the Einstellung Effect.** As corroborated by our failure mode analysis (Figure 1), highly capable models like GPT-5 exhibit a disproportionately high rate of *Blindness*.This suggests that as stronger models so effectively capture the prior distribution of training data, they become overly anchored to statistical shortcuts, literally filter out low-probability counter-evidence during perception. This empirical evidence solidifies our "Stronger Priors, Stronger Blindness" phenomenon, demonstrating that reasoning robustness cannot emerge from probability optimization alone.

2. Existing Agents Amplify Cognitive Bias. Compared to the base model (Qwen3-32B), the multi-agent framework MDAgent does not yield the expected improvements and even exhibits degradation. We attribute this to two factors: (1) **Noise Amplification:** The significant drop in $\boldsymbol{Acc_{base}}$ (40.26% $\rightarrow$ 29.70%) suggests that without causal constraints, the diverse viewpoints introduced by multi-agent debate act as noise rather than signal. (2) **Bias Amplification:** The stagnation in $\boldsymbol{Acc_{rob}}$ and high $\boldsymbol{R_{bias}}$ indicate that the "debate" mechanism, when faced with strong Einstellung traps, devolves into **Consensus Bias**, reinforcing the incorrect intuitive consensus rather than correcting it.

3. Effectiveness of Evidence-Based Architecture. In contrast, ECR-Agent (based on Qwen3-32B) achieves a qualitative leap in performance. It significantly boosts fundamental capability ($\boldsymbol{Acc_{base}} \rightarrow 69.49\%$) while doubling robustness ($\boldsymbol{Acc_{rob}} \rightarrow 24.21\%$) and reducing the bias rate ($\boldsymbol{R_{bias}} \rightarrow 33.75\%$). This demonstrates that introducing **Structural Causal Reasoning** and **Evidence Audit** mechanisms is key to breaking the Einstellung Effect. Unlike baselines that rely on internal parametric memory, ECR-Agent enforces an evidence-based reasoning process that prioritizes "evidence" over "probability," effectively circumventing the Einstellung Traps.

D.3 Impact of Scale and Pathology

Scaling Ineffectiveness. Figure 5 visualizes the relationship between $\boldsymbol{R_{bias}}$ and $\boldsymbol{Acc_{base}}$. The results show no significant linear negative correlation, with data points widely scattered. Frontier models like GPT-5, despite possessing extreme fundamental capability (right side of X-axis), still exhibit very high bias rates (top of Y-axis). This indicates that reasoning robustness does not emerge naturally from scale. Without a structured verification mechanism, even advanced CoT reasoning remains susceptible to being trapped in the Einstellung Effect by strong statistical priors.

Pathology-Dependent Vulnerability. The clustering patterns in Figure 5 and the heatmap in Figure 4 reveal the structural nature of the Einstellung Effect:

- **Clustering:** Pathologies like *Pneumonia* and *Pericarditis* consistently appear in the High Bias Cluster across almost all models. This reveals strong **Spurious Correlations** in the training data.

- **Variance:** Conversely, pathologies like *Influenza* show high variance, suggesting that when statistical priors are weaker, some models can successfully reason through distractors.

This pathology dependence confirms the systemic vulnerability of probabilistic models when facing "High-Confidence Prior vs. Low-Confidence Evidence" conflicts. ECR-Agent succeeds by transforming the "probability prediction problem" into an "evidence verification problem" via Causal Intervention, structurally blocking the propagation of spurious correlations.

D.4 Cross-Domain Generalizability

To validate the generalizability of our framework beyond the counterfactual scenarios of MedEinst, we evaluate ECR-Agent on MedChain (Liu et al., 2024a), a real-world clinical benchmark. As shown in Table A1, ECR-Agent outperforms standard multi-agent debate frameworks (e.g., MDAgent, DyLAN) by a significant margin. While the domain-specific MedChain-Agent achieves slightly higher performance (48.07%), ECR-Agent yields a competitive 46.09% accuracy relying solely on its generalized **Dynamic Causal Inference (DCI)** pipeline without domain-specific adaptation, proving its broad clinical applicability.

D.5 Robustness Stress Test with Explicit Warning

To rule out the possibility that the Einstellung Effect is an artifact of insufficient prompting, we design a stress test by injecting an explicit warning into the Zero-shot CoT prompt for Qwen3-32B and Qwen3-235B:

"CRITICAL WARNING: This case may contain misleading distractors (red herrings) that mimic common diseases. You

Multi-Agent Framework	Diagnosis Accuracy
DyLAN	38.63%
MDAgent	39.73%
MedAgent	41.02%
MDAgent + RAG	41.98%
MedChain-Agent	48.07%
ECR-Agent (Ours)	**46.09%**

Table A1: Diagnosis Accuracy on MedChain Benchmark.

must strictly base your diagnosis on specific discriminative evidence, not just general patterns or prior probabilities. Be skeptical of 'obvious' patterns if key evidence contradicts them." As Table A2 demonstrates, the explicit warning fails to mitigate the bias. Instead, it pushes the models to fit general patterns more aggressively, leading to higher Baseline Accuracy but significantly higher Bias Trap Rates and lower Robust Accuracy. This structural failure reinforces the necessity of our **Dynamic Causal Graph Reasoning (DCGR)** module, which structurally enforces evidence verification, proving that mitigating the Einstellung Effect requires a fundamental shift from probabilistic pattern matching to rigorous causal grounding.

D.6 Validation of the Bias Trap Rate Metric

To verify that our primary metric, the Bias Trap Rate ($\boldsymbol{R_{bias}}$), accurately captures the **cognitive rigidity** inherent in the Einstellung Effect rather than random stochastic noise or general performance degradation, we conducted two supplementary analyses on the error distributions.

Directionality of Errors (Rigidity Ratio): We investigated whether model failures on Trap Cases ($\boldsymbol{x^t}$) are random or systematically biased toward the prior. We distinguish between *Rigid Reversion*—where the model ignores discriminative evidence and reverts to the control label $\boldsymbol{y_{gt}}$—and *Stochastic Errors* (predictions of unrelated pathologies). We define the **Rigidity Ratio** as:

$$\text{Rigidity Ratio} = \frac{\text{Count(Rigid Reversion)}}{\text{Count(Rigid Reversion)} + \text{Count(Stochastic Errors)}} \quad (6)$$

As shown in Table A3, failures are overwhelmingly directional. For instance, 88.24% of Gemini-2.5-Pro's errors involve a strict reversion to the initial prior. This high ratio confirms that the misdiagnoses are not random "brittleness" but are driven by an active pull toward the **local attractor** established by the initial symptoms.

Rank Rigidity and Overconfidence: We further analyzed the models' internal ranking and confidence when encountering Einstellung traps. As shown in Table A4, despite the presence of explicit counter-evidence in $\boldsymbol{x^t}$, the prior class $\boldsymbol{y_{gt}}$ experienced a negligible **Ranking Drop** ($\boldsymbol{\Delta}$ Rank $\boldsymbol{< 1.0}$), while the models maintained an average **Confidence** of $\boldsymbol{> 85\%}$ in their incorrect predictions. This combination of structural rank rigidity and overconfidence empirically supports our "Stronger Priors, Stronger Blindness" thesis (§5.5), proving that $\boldsymbol{R_{bias}}$ effectively identifies cases where statistical shortcuts override patient-specific evidence.

D.7 Mathematical Formulation and Raw Count Breakdown

To provide a transparent view of the evaluation space and explicitly connect our pair-level logical conditions to the aggregate metrics, we mathematically formalize the mutually exclusive outcomes. Given a counterfactual pair in the competent subset $\boldsymbol{S_{correct\_control}}$ (i.e., instances where the model successfully diagnoses the control case, $\boldsymbol{f(x^c) = y_{gt}}$), the prediction on the corresponding trap case $\boldsymbol{f(x^t)}$ must strictly fall into one of three distinct categories:

1. **Robust Success ($\boldsymbol{f(x^t) = y_{bias}}$):** The model successfully overrides its initial prior using the new discriminative evidence and correctly diagnoses the trap case.
2. **Rigid Reversion / Einstellung Effect ($\boldsymbol{f(x^t) = y_{gt}}$):** The model relies on statistical shortcuts and falls into the bias trap, rigidly adhering to the prior diagnosis despite contradictory evidence.
3. **Stochastic Error ($\boldsymbol{f(x^t) \notin \{y_{gt}, y_{bias}\}}$):** The model predicts an unrelated pathology, indicating a general reasoning failure or calibration error rather than a directional cognitive bias.

Table A5 presents the granular raw count breakdown for three frontier models. This transparent view mathematically substantiates our core thesis: the elevated error rates observed in frontier models are not artifacts of general incompetence (Stochastic Errors). Instead, the overwhelming majority of failures are strictly confined to the specific Einstellung pattern ($\boldsymbol{y_{gt}}$). For example, out of 2,882 competent control predictions made by Gemini-2.5-Pro, an absolute majority of 1,755 cases (60.9%) resulted in a strict reversion to the prior.

E Case Study

To demonstrate the efficacy of MEDEINST in benchmarking the Einstellung Effect and the robustness of ECR-AGENT, we present a detailed analysis of **Case 100473**. This case represents a high-stakes emergency scenario where the baseline model succumbed to a "Pattern Matching" trap, while our agent successfully corrected the diagnosis through causal graph reasoning.

E.1 Case Overview

- **Ground Truth:** Pulmonary Embolism (PE).
- **Trap Type:** *Distractor Injection* (Family History of Pneumothorax) + *Evidence Substitution* (History of DVT).
- **Baseline Intuition:** Spontaneous Pneumothorax.
- **ECR-Agent Verdict:** Overturn → Pulmonary Embolism.

E.2 Narrative Comparison

Table A6 illustrates the minimal yet critical differences between the Control and Trap cases. The Trap case introduces a strong "Red Herring" (Family History) while subtly embedding the key discriminative evidence (DVT History).

E.3 Reasoning Trace Analysis

The baseline model (intuition) anchored on the "Young Male + Sudden Chest Pain + Family History" pattern, incorrectly diagnosing Spontaneous Pneumothorax. Below is the **reconstructed audit log** from the ECR-AGENT's *Evidence Audit*, demonstrating how it utilized the Causal Graph to overturn this error.

Model	Setting	Baseline Acc (↑)	Robust Acc (↑)	Bias Trap Rate (↓)
Qwen3-32B	Vanilla Zero-shot CoT	40.25%	11.86%	43.46%
Qwen3-32B	**+ Explicit Warning**	52.60%	8.70%	**60.50%**
Qwen3-235B	Vanilla Zero-shot CoT	40.51%	11.40%	43.32%
Qwen3-235B	**+ Explicit Warning**	53.30%	9.20%	**58.00%**

Table A2: Stress Test Results with Explicit Warning Prompt.

Model	Rigidity Ratio (↑)
Gemini-2.5-Pro	88.24%
GPT-5	85.27%
Kimi-k2	83.05%
Qwen3-32B	79.65%

Table A3: Rigidity Ratio across representative models, indicating the proportion of errors that revert to the prior label y_{gt}.

Model	Ranking Drop (↓)	Confidence in y_{gt} (↑)
Qwen3-32B	0.89	85.4%
Qwen3-235B	0.94	85.2%

Table A4: Internal ranking and confidence analysis on MedEinst trap cases.

Evidence Audit Log:
Initial Hypothesis: Spontaneous Pneumothorax (Probability: High).
Tier 1: Fatal Conflict Check
- *Check:* Pneumothorax typically requires specific imaging confirmation (e.g., "Spontaneous pneumothorax on imaging").
- *Result:* Missing/Shadow Node. The narrative does not mention imaging evidence.
Tier 2: Pivot Competition
- *Pivot 1 (DVT History):* Strongly supports **Pulmonary Embolism** (Risk Factor). Matched in Patient Narrative.
- *Pivot 2 (Family History):* Supports Pneumothorax, but is a weak "General" feature compared to the specific risk factor of DVT.
- *Pivot 3 (Hypoxia + Tachycardia):* Supports both, but biologically more severe in PE.
Decision: The presence of "History of DVT" is a distinct Pivot Node that rules out Pneumothorax (as a primary cause) and strongly supports PE. The initial intuition was biased by the family history.
Final Verdict: OVERTURN → Pulmonary Embolism.

Model	Total ($\lvert S_{correct\_control}\rvert$)	Robust Success	Rigid Reversion (EE)	Stochastic Error
GPT-5	2,921	849	**1,515 (51.9%)**	557 (19.1%)
Gemini-2.5-Pro	2,882	590	**1,755 (60.9%)**	537 (18.6%)
DeepSeek-R1	2,270	609	**1,047 (46.1%)**	614 (27.0%)

Table A5: Raw count breakdown of outcomes on the trap cases for frontier models. The percentages in the "Rigid Reversion" column correspond exactly to the Bias Trap Rate (R_{bias}) reported in the main text.

E.4 Interpretability: Evidence Balance Sheet

The core of ECR-AGENT's interpretability lies in its explicit **Causal Graph**. Table A7 details the "Evidence Balance Sheet" for Case 100473.

The agent constructs a graph connecting the Patient Observations (P_{obs}) to the Knowledge Nodes (K_{nodes}) of competing diagnoses. The decision is driven by **Pivot Nodes**—features that logically distinguish between the two conditions.

F Extended Discussion: Theoretical Grounding and Comparative Analysis

While the main text outlines the broad landscape of medical LLMs, this appendix provides a deeper theoretical analysis of why existing paradigms—specifically Multi-Agent Collaboration and Retrieval-Augmented Generation (RAG)—insufficiently address the Einstellung Effect, and how our ECR-AGENT fundamentally differs by aligning with Causal Inference theories.

F.1 Verification vs. Consensus: The Limits of Multi-Agent Debate

Recent agentic frameworks like MDAgents (Kim et al., 2024) and MedAgents (Tang et al., 2024) rely on "collaboration" or "debate" strategies, assuming that diverse personas will cancel out individual errors. However, this assumption holds only when errors are independent and randomly distributed.

In the context of the **Einstellung Effect**, errors are not random but *systematic*. As shown in our experiments (Table 1), strong statistical priors act as a "common distractor" that misleads the majority of models/agents similarly.

- **Consensus Bias:** When the "intuitive but wrong" diagnosis is statistically dominant, multi-agent debate often devolves into **Consensus Bias** (Schmidgall et al., 2024a,b). Agents tend to converge on the most likely probabilistic token rather than the ground truth evidence.
- **Our Solution (Veto by Evidence):** Unlike debate frameworks that optimize for *agreement*, ECR-AGENT optimizes for *falsification*. By introducing **Pivot Nodes** (Section 4.2.2), our agent grants a single piece of discriminative evidence the power to "veto" the majority consensus, mirroring the clinical principle that "one proven contradiction outweighs a thousand probabilities".

F.2 Dynamic Inference vs. Static Knowledge: The Limits of RAG

Retrieval-Augmented Generation (RAG) systems, such as MedGraphRAG (Wu et al., 2024) and PrimeKG (Chandak et al., 2023), attempt to mitigate hallucinations by retrieving external knowledge. While effective for factual queries, standard RAG faces structural limitations in **Counterfactual Differential Diagnosis**:

Feature	Control Case (x^c)	Trap Case (x^t)
Demographics	Male, 22 years old	Male, 22 years old
Chief Complaint	Sudden "knife-like" chest pain, Dyspnea	Sudden "knife-like" chest pain, Dyspnea
Key Evidence	History of **Spontaneous Pneumothorax**	History of **Deep Vein Thrombosis (DVT)**
Distractors	Family history of Pneumothorax	Family history of Pneumothorax
Associated Sx	Tachycardia, Hypoxia	Tachycardia, Hypoxia
Model Diagnosis	Spontaneous Pneumothorax (✓)	Spontaneous Pneumothorax (**Error**)

Table A6: Comparison of the Control and Trap narratives. The **Trap Case** replaces the patient's personal history with DVT (a risk factor for PE) but retains the family history of Pneumothorax, triggering the Einstellung Effect in baseline models.

Diagnosis Candidate	Node Type	Relation to Patient	Clinical Feature Content
Diagnosis A: Pulmonary Embolism (Correct)			
	Pivot	**Match**	**History of Deep Vein Thrombosis (DVT)**
	Pivot	**Match**	Sudden onset dyspnea with tachycardia & hypoxia
	Pivot	**Match**	Sharp pleuritic chest pain exacerbated by inspiration
	General	Match	Dyspnea with sudden onset
Diagnosis B: Spontaneous Pneumothorax (Intuition/Error)			
	Pivot	**Missing**	Spontaneous pneumothorax on imaging
	General	Match	History of spontaneous pneumothorax & family history
	Pivot	Match	Acute onset pleuritic chest pain
	Pivot	**Rule Out**	*Presence of prior Deep Vein Thrombosis (DVT)*

Table A7: **Evidence Balance Sheet.** The table shows why the agent favored PE over Pneumothorax. While Pneumothorax has matching symptoms (chest pain), it lacks its critical Pivot evidence (Imaging) and is actively ruled out by the presence of DVT, which is a Pivot Match for PE.

- **Static vs. Dynamic:** RAG retrieves *static* associations (e.g., "Pulmonary Embolism causes Chest Pain") but lacks the mechanism to construct a *patient-specific* causal graph. It cannot dynamically evaluate "What if this specific symptom was absent?" or "Why is this overlapping symptom non-discriminative in this specific context?".

- **Associative vs. Causal:** RAG fundamentally enhances *Associative Reasoning* (Pearl's Layer 1) by adding more context to the prompt. It does not perform *Intervention* (Layer 2).

- **Our Solution:** ECR-AGENT does not just retrieve knowledge; it structures it into a **Dynamic Causal Graph**. By explicitly modeling *Match*, *Conflict*, and *Shadow* relations, we transform static knowledge into active reasoning tools that can perform logical interventions on the patient's narrative.

F.3 Theoretical Grounding: Mapping Diagnosis to the Causal Hierarchy

Our framework is theoretically grounded in the integration of Evidence-Based Medicine (EBM) (Sackett, 1997) with Pearl's Causal Hierarchy (Pearl and Mackenzie, 2018). We provide a formal mapping of these cognitive processes:

1. **Layer 1: Association.** *Clinical Equivalent:* Pattern Recognition / Intuition. *Implementation:* Our **Dual-Pathway Perception** module generates initial hypotheses based on $\boldsymbol{P(Diagnosis|Symptoms)}$. This is where the Einstellung Effect (statistical bias) originates. Crucially, current scaling paradigms primarily optimize this associative layer. While scaling laws ensure that larger models approximate $\boldsymbol{P(Diagnosis|Symptoms)}$ with increasing precision, they remain mathematically confined to Layer 1. This theoretical bottleneck perfectly explains our "Stronger Priors, Stronger Blindness" observation: as scaling pushes associative pattern matching to its theoretical limit, the resulting rigid priors render the model incapable of spontaneously leaping to Layer 2 (Intervention) to falsify its initial hypotheses.

2. **Layer 2: Intervention.** *Clinical Equivalent:* Differential Diagnosis / Testing. *Implementation:* Our **Forward Causal Reasoning** simulates the act of "intervening" to find truth. We define **Pivot Nodes** as the minimal intervention set $\boldsymbol{do(X)}$ required to distinguish between competing hypotheses $\boldsymbol{d_i}$ and $\boldsymbol{d_j}$. This aligns with Richens et al. (2020), who proved that optimal diagnosis requires maximizing the Information Gain of interventions.

3. **Layer 3: Counterfactuals.** *Clinical Equivalent:* Diagnostic Verification / Audit. *Implementation:* Our **Backward Causal Reasoning** and **Evidence Audit** perform the counterfactual check: "Given diagnosis $\boldsymbol{d}$, what symptom $\boldsymbol{s}$ *would have been* observed?". The detection of **Shadow Nodes** (missing expected evidence) formally represents the violation of counterfactual expectations ($\boldsymbol{P(s_{missing}|do(d)) \approx 0}$), allowing the model to reject high-probability but causally inconsistent traps.

This rigorous mapping demonstrates that ECR-AGENT is not merely an engineering improvement but a step towards **Causal AI** in medicine, moving beyond the *Curve Fitting* limitations of standard LLMs (Richens et al., 2020).

G Future Work: Addressing Complex Clinical Realities

As discussed in the main text's Limitations section, real-world clinical practice often involves complexities that extend beyond acute, single-disease presentations. Based on the foundational architecture of **ECR-Agent**, we outline three key avenues for future methodological extension:

- **Rare Diseases (Mitigating Weak Priors):** Current LLMs lack sufficient pre-training priors for rare pathologies. While ECR-Agent's Critic-Driven Graph & Memory Evolution (CGME) is highly data-efficient (requiring only 853 training samples to initialize illness graphs in our experiments), future iterations will focus on dynamically retrieving and constructing causal subgraphs directly from minimal clinical literature (e.g., sparse case reports) to compensate for this scarcity of statistical priors.

- **Multimorbidity and Graph Composability:** The current Dynamic Causal Inference evaluates candidates as mutually exclusive, which struggles with co-infections or concurrent conditions. However, because our CGME module constructs independent, **pathology-centric illness graphs**, the framework possesses inherent *structural composability*. Future work will leverage this property: when multiple **Pivot Nodes** from distinct pathologies are simultaneously verified, their respective illness graphs can be naturally merged into a unified multimorbidity causal network, smoothly shifting the paradigm from single-disease classification to complex co-morbidity modeling.

- **Longitudinal Scenarios (Temporal Tracking):** MedEinst currently evaluates isolated, cross-sectional patient narratives. Because the temporal evolution of symptoms is critical in longitudinal care, our proposed *Shadow Nodes*—which currently penalize missing expected cross-sectional evidence—can naturally be repurposed as longitudinal anchors. This adaptation will allow the agent to generate dynamic "follow-up checklists" for subsequent patient visits, extending causal verification across time.

H Data Samples

To demonstrate the realistic clinical presentation of MEDEINST, Figure 8 displays the raw input narratives for Case 100473 as they appear to the model.

We adopt the structured format from the DDXPlus dataset, which organizes clinical observations into `Symptoms` and `Antecedents` with hierarchical indentation. The figure highlights the **counterfactual intervention**: while the lengthy symptom description and the "Family history" distractor remain identical, the specific patient history in the `Antecedents` section is surgically altered from "Spontaneous pneumothorax" (Control) to "Deep vein thrombosis" (Trap).

I Prompts Details

To ensure the reproducibility of our work, we provide the full system prompts used in both the MEDEINST benchmark construction pipeline and the ECR-AGENT reasoning framework.

I.1 MedEinst Benchmark Construction

Tables A8, A9, A10, and A11 detail the prompts for the four-stage adversarial data construction pipeline.

I.2 ECR-Agent Reasoning Framework

Tables A12, A13, and A14 detail the prompts for the three-phase causal reasoning engine.

Control Case (x^c): Spontaneous Pneumothorax

```
Sex: M, Age: 22
Geographical region: North America

Symptoms:
---------
- I feel pain.
  - The pain is:
    * heartbreaking
    * a knife stroke
  - The pain locations are:
    * side of the chest(R)
    * breast(R)
    * breast(L)
- On a scale of 0-10,
  the pain intensity is 6
- The pain radiates to
  these locations:
  * nowhere
- On a scale of 0-10,
  the location precision is 2
- On a scale of 0-10,
  the speed of onset is 9
- I am experiencing shortness
  of breath or
  difficulty breathing in a significant way.
- I have pain that is increased
  when I breathe
  in deeply.
- I have tachycardia.
- I have hypoxia.

Antecedents:
------------
- I have had a spontaneous pneumothorax.
- I smoke cigarettes.
- One or more of my family members have had
  a pneumothorax.
- I have not traveled out of the country in
  the last 4 weeks.
```

Trap Case (x^t): Pulmonary Embolism

```
Sex: M, Age: 22
Geographical region: North America

Symptoms:
---------
- I feel pain.
  - The pain is:
    * heartbreaking
    * a knife stroke
  - The pain locations are:
    * side of the chest(R)
    * breast(R)
    * breast(L)
- On a scale of 0-10,
  the pain intensity is 6
- The pain radiates to
  these locations:
  * nowhere
- On a scale of 0-10,
  the location precision is 2
- On a scale of 0-10,
  the speed of onset is 9
- I am experiencing shortness
  of breath or
  difficulty breathing in a significant way.
- I have pain that is increased
  when I breathe
  in deeply.
- I have tachycardia.
- I have hypoxia.

Antecedents:
------------
- I have had a deep vein thrombosis (DVT). <!!>
- I smoke cigarettes.
- One or more of my family members have had
  a pneumothorax.
- I have not traveled out of the country in
  the last 4 weeks.
```

Figure 8: Side-by-side comparison of the raw clinical narratives for Case 100473. The text is presented in the original DDXPlus format used as input for the LLMs. The **Trap Case** (Right) contains a minimal edit in the `Antecedents` section (marked with <!!>), replacing the history of pneumothorax with DVT, while retaining the misleading family history.

Discriminative Feature Extraction Prompt (Step 1)

`System Prompt:` You are a senior medical expert. Your task is to perform a differential diagnosis based on the provided reference knowledge.
CRITICAL INSTRUCTION: You MUST use the 'Reference Knowledge' below as your ONLY source of truth. Do not use your own internal knowledge.
Reference Knowledge:
- Typical Symptoms of 'distractor_disease': distractor_symptoms
- Typical Symptoms of 'truth_disease': truth_symptoms
Patient Narrative: — control_narrative —
Your Task:
1. Compare the patient's narrative with the two lists of typical symptoms.
2. Identify the SINGLE MOST CRITICAL and ATOMIC phrase within the narrative that is a typical symptom of 'distractor_disease' but NOT a typical symptom of 'truth_disease'.
Response Format: If you successfully identify such a phrase, respond in JSON format with one key: {"narrative_A": "the exact phrase from the narrative"}.

Table A8: Prompt used to extract the key discriminative evidence ($\boldsymbol{k_{gt}}$) that supports the control diagnosis.

Trap Information Generation Prompt (Step 2)

`System Prompt:` You are a medical writer. Your task is to generate an 'isomorphic' clinical finding, grounded in the provided reference knowledge.
CRITICAL INSTRUCTION: You MUST choose an evidence from the 'Reference Knowledge for truth_disease' list below.
Reference Knowledge for truth_disease: ...
The original phrase (pointing to distractor_disease): "narrative_A"
Your Task:
1. Review the list of typical evidence for 'truth_disease'.
2. Select one evidence from those lists that is most 'isomorphic' to the original phrase in terms of clinical gravity and role.
3. Rephrase the selected evidence into a concise, ATOMIC phrase a patient would say.
Respond in JSON format with one key: "narrative_B".

Table A9: Prompt used to generate the misleading trap feature ($\boldsymbol{k_{trap}}$) based on the bias disease knowledge.

Differential Features Rewrite Prompt (Step 3)

`System Prompt:` You are an expert medical writer performing a "Cognitive Surgery". Your task is to seamlessly rewrite a clinical narrative to change its diagnostic direction.
Original Patient Narrative: — control_narrative —
Your instructions:
1. Locate the phrase "narrative_A" in the text.
2. Rewrite the single sentence containing this phrase to instead convey the new core information: "narrative_B".
3. **Crucially, you MUST ensure the new sentence is grammatically perfect and logically fits the surrounding context.** The style and tone must match the original exactly.
4. **Do not change any other part of the narrative.**
Respond with only the complete, rewritten patient narrative text.

Table A10: Prompt used to inject the trap feature into the patient narrative ($\boldsymbol{x^c \rightarrow x^t}$).

Inter-Model Verification Prompt (Step 4)

`System Prompt:` You are an expert clinical diagnostician and medical narrative analyst, acting as an impartial judge. Your task is to assess the quality of a synthetically modified "Trap Case".
Your evaluation must be based on three precise criteria:
1. **Diagnostic Correctness (Boolean):** When reading the trap case narrative independently, does it provide sufficient evidence to make its stated ground truth ('trap_gt') a plausible and likely diagnosis?
2. **Medical Plausibility (1-10 Scale):** How medically believable is the complete trap case narrative?
3. **Narrative Fluency (1-10 Scale):** How well was the new information integrated?
...

Table A11: LLM-as-a-Judge prompt for verifying the quality and validity of generated trap cases.

Analytic Problem Representation Prompt (Dual-Pathway Perception)

`System Prompt:` You are a Senior Clinical Diagnostician and Expert Medical Scribe. Your task is to perform "Problem Representation" on a raw patient case.
OBJECTIVE: Transform the patient's raw narrative into a structured list of **P-Nodes (Patient Features)** using precise **Medical Semantic Qualifiers**.
THE PROCESS:
1. Translate Time ... 2. Translate Symptoms ... 3. Filter ... 4. Synthesize ...
OUTPUT SCHEMA (JSON):
Return a single JSON object with two keys: 'problem_representation_one_liner' and 'p_nodes' (containing id, content, original_text, status).
RULES: Only mark status: "Absent" if the text explicitly says "no", "denies", or "without".

Table A12: Prompt for extracting structured patient observations ($\boldsymbol{P_{obs}}$) from raw text.

Pivot Node Discovery Prompt (Causal Graph Reasoning)

`System Prompt:` You are an Expert Diagnostician performing a comprehensive Differential Diagnosis.
Step 1: Disease-by-Disease Analysis ...
Step 2: Cross-Disease Comparison (Matrix Analysis)
Create a mental discrimination matrix: Which features are UNIQUE to one disease? Which features RULE OUT certain diseases?
OUTPUT JSON SCHEMA:
You MUST output a JSON object with: "k_nodes": [{ "content": "...", "type": "Pivot", "importance": "Pathognomonic", "supported_candidates": [...], "ruled_out_candidates": [...] }]
FIELD DEFINITIONS:
- **Pivot**: Discriminating feature that helps distinguish between 2+ diseases.

Table A13: Prompt for identifying Pivot Nodes to differentiate between competing hypotheses.

Evidence Audit & Final Decision Prompt (Evidence Audit)

`System Prompt:` You are the **Chief Medical Auditor** and Final Decision Maker. Your goal is to audit the reasoning of a "System 1" (Initial Intuition) agent using a "System 2" (Causal Graph) evidence map.
THE LOGIC HIERARCHY (Follow Strictly)
Tier 1: The Safety Sentinel (Fatal Conflicts)
Rule: If a Candidate requires a symptom that is 'Essential', but the Patient explicitly has 'Status: Absent', then this Candidate is DISQUALIFIED.
Tier 2: The Pivot Competition (Differential Diagnosis)
Rule: A Candidate supported by a **matched Pivot Feature** is superior to a Candidate supported only by General features.
Tier 3: The Shadow & Coverage Audit (Tie-Breaker)
Select the candidate with the highest explanatory coverage and fewest unexplained conflicts.

Table A14: Prompt for the final evidence audit, applying the Tiered Logic Hierarchy to select the diagnosis.