arXiv:2509.11078v2 [cs.AI] 6 Jan 2026

Patient-Zero: Scaling Synthetic Patient Agents to Real-World Distributions without Real Patient Data

Yunghwei Lai[1] **Ziyue Wang**[1] **Weizhi Ma**[2*] **Yang Liu**[1,2*]
[1]Dept. of Comp. Sci. & Tech., Institute for AI, Tsinghua University
[2]Institute for AI Industry Research (AIR), Tsinghua University

Abstract

Synthetic data generation with Large Language Models (LLMs) has emerged as a promising solution in the medical domain to mitigate data scarcity and privacy constraints. However, existing approaches remain constrained by their derivative nature, relying on real-world records, which pose privacy risks and distribution biases. Furthermore, current patient agents face the *Stability-Plasticity Dilemma*, struggling to maintain clinical consistency during dynamic inquiries. To address these challenges, we introduce **Patient-Zero**, a novel framework for *ab initio* patient simulation that requires *no real medical records*. Our Medically-Aligned Hierarchical Synthesis framework generates comprehensive and diverse patient records from abstract clinical guidelines via stratified attribute permutation. To support rigorous clinical interaction, we design a Dual-Track Cognitive Memory System to enable agents dynamically update memory while preserving logical consistency and persona adherence. Extensive evaluations show that **Patient-Zero** establishes a new state-of-the-art in both data quality and interaction fidelity. In human expert evaluations, senior licensed physicians judge our synthetic data to be ***statistically indistinguishable from real human-authored data*** and higher in clinical quality. Furthermore, downstream medical reasoning model trained on our synthetic dataset shows substantial performance gains (MedQA +24.0%; MMLU +14.5%), demonstrating the practical utility of our framework.

1 Introduction

Large Language Models (LLMs) have shown remarkable capabilities in generative tasks (OpenAI, 2024; Zhang et al., 2025), increasingly being used for synthetic data construction across various domains (Li et al., 2023b; Guo and Chen, 2024; Nahid and Hasan, 2024; Karst et al., 2024). In the medical field, high-quality synthetic data offers a promising solution for high costs and strict privacy constraints in real-world patient data (Goncalves et al., 2020a). Dominant approaches for patient record generation span retrieval-reasoning (Xu et al., 2025), reinforcement learning (RL) (Das et al., 2024), generative adversarial network (GAN) (Gwon et al., 2024; Li et al., 2023a), and alternative methods (Sun et al., 2024; Guillaudeux et al., 2023; Tornqvist et al., 2024; Liu et al., 2024; Walonoski et al., 2017).

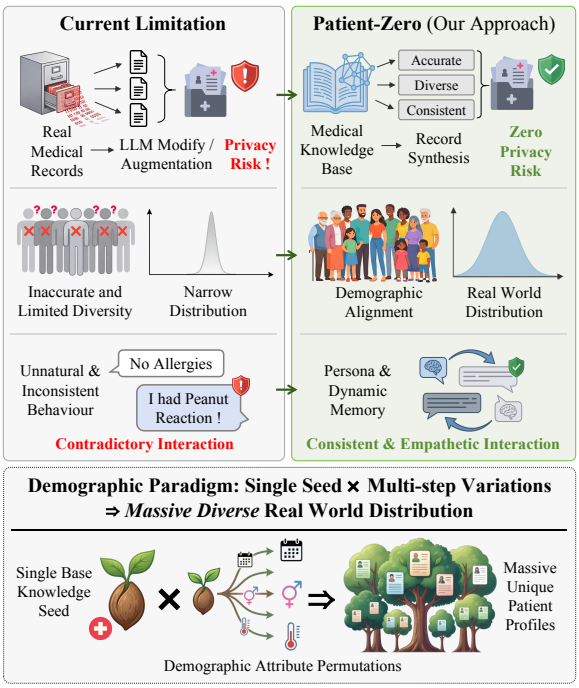


Figure 1: Our **Patient-Zero** Paradigm. While conventional methods are constrained by the derivative nature of real-world data, such as privacy risks and distribution biases, our framework enables *ab initio* patient simulation. Instead of using sensitive medical records as seed, **Patient-Zero** constructs patient agents *from scratch* using medical knowledge, achieving zero privacy risk while maintaining clinical consistency throughout synthetic data generation and interactive simulation.

However, existing approaches are fundamentally constrained by two primary limitations. First, regarding **data quality**, methods derived from *real-*

*Correspondence to Weizhi Ma (mawz@tsinghua.edu.cn), Yang Liu (liuyang2011@tsinghua.edu.cn)

world medical records face inherent bottlenecks: **privacy risks**, **medical inconsistency**, and **limited diversity**. Even with de-identification methods to exclude personal identifiers, generative approaches may still pose privacy risks (Chen and Esmaeilzadeh, 2024) or fail to strictly adhere to clinical guidelines (Chen et al., 2025). Furthermore, the reliance on specific source datasets restricts the distribution, failing to capture the full complexity of real-world scenarios (Giuffrè and Shung, 2023). Second, in terms of **interactive capability**, existing patient agents often suffer from unnatural and inconsistent behaviors in clinical inquiries (Graf et al., 2024; Cook et al., 2025), or are confined to specific domains (Wang et al., 2024; Louie et al., 2024; Lee et al., 2025), rendering them unable to generalize across multi-specialty clinical scenarios. This motivates a fundamental research question:

Can we generate realistic interactive patient agents without exposing any real-world records?

In this work, we propose **Patient-Zero** as an effective solution to this challenge, which is a novel hierarchical generation framework that synthesizes medically-aligned patient records without relying on any real patient data. It employs a hierarchical process anchored in clinical guidelines, leveraging *epidemiological attribute permutations* to evolve abstract disease concepts into granular patient records. To bridge the gap between static records and dynamic interaction, we introduce a Natural Language Inference Verifier (*NLI-Verifier*), ensuring both logical consistency and persona adherence for robust memory updates. To validate the practical utility of this framework, manual evaluations by licensed physicians confirm that **Patient-Zero** produces narratives that strictly adhere to clinical guidelines and professional standards. Our key contributions are summarized as follows:

- **(Framework)** We introduce **Patient-Zero**, a novel framework that generates large-scale synthetic patient records *ab initio*, without utilizing any real patient data, ensuring strict clinical adherence while achieving scalable data diversity that aligns with real-world distributions.
- **(Simulation)** We propose a *Dual-Track Cognitive Memory System* to resolve interaction inconsistencies. Patient agents can perform memory updates with logical consistency and persona adherence during extended clinical dialogues.
- **(Validation)** Extensive evaluations demonstrate that **Patient-Zero** establishes new state-of-the-art results in synthetic data quality and interaction fidelity, while yielding substantial downstream performance gains. In expert evaluation, senior licensed physicians judged our synthetic data to be statistically indistinguishable from real human-authored data and higher in clinical quality.

2 Related Work

Synthetic Medical Record Generation In the healthcare sector, strict privacy regulations have significantly restricted access to real Electronic Health Records (EHRs), driving the research into synthetic data generation (Kruse et al., 2017; Goncalves et al., 2020b). Early generative approaches predominantly utilized GAN (Macedo et al., 2024; Feng et al., 2024; Lange et al., 2024). With the paradigm shift toward LLM, models are conditioned on seed data retrieved from real-world records, including methods such as retrieval-reasoning (Xu et al., 2025), RL (Das et al., 2024), and other approaches (Sun et al., 2024; Yu et al., 2025; Liu et al., 2024; Guillaudeux et al., 2023; Tornqvist et al., 2024; Walonoski et al., 2017). However, they remain fundamentally dependent on real-world patient data, which poses privacy risks and limits the diversity to the distribution of the source datasets. In contrast, **Patient-Zero** generates patient records *without real private data* while *aligning with diverse real-world distribution*.

Patient Agent Simulation Simulated patient agents serve as a scalable environment for training medical professionals and evaluating diagnostic reasoning (Lizée et al., 2024; Zhu et al., 2024). Current methodologies generally fall into two categories. The first prioritizes behavioral and psychological fidelity, such as cognitive models and role-playing scenarios (Wang et al., 2024; Louie et al., 2024; Lee et al., 2025; Du et al., 2024; Wasenmüller et al., 2024), while the second emphasizes clinical accuracy and reasoning reliability, ensuring the agent provides factually correct and logically sound responses (Yu et al., 2025; Li et al., 2024). Nevertheless, strictly scripted agents often appear robotic, whereas open-ended generative models tend to forget their initial clinical persona or contradict themselves during extended interactions. **Patient-Zero** address this by implementing a real-time memory verification mechanism, maintaining strict logical consistency and persona adherence while preserving natural dialogue flexibility.

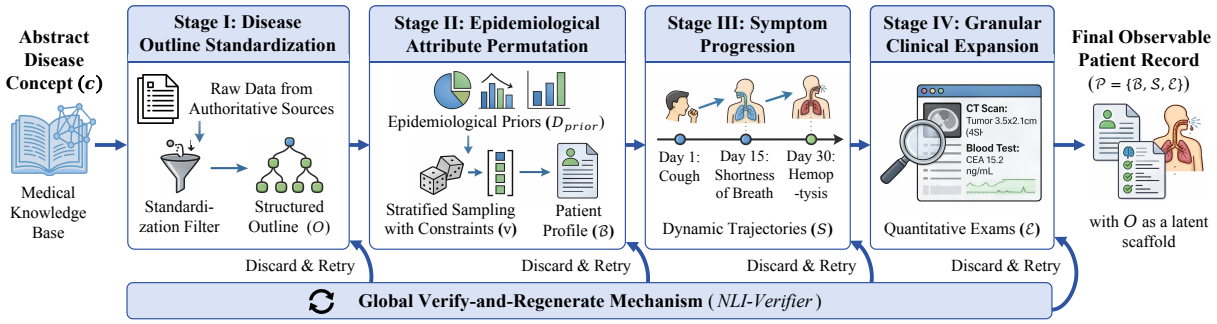


Figure 2: Overview of our **Patient-Zero** Hierarchical Synthesis Framework. The pipeline factorizes patient generation into a four-stage causal chain ($c \to \mathcal{O} \to \mathcal{B} \to \mathcal{S} \to \mathcal{E}$). Starting from an abstract disease concept (c), the system progressively expands medical details: I) Standardizing noisy knowledge into a structured outline; II) Sampling epidemiological attributes via constrained permutation; III) Evolving dynamic symptom trajectories; and IV) Generating granular, quantitative examination results. A global *verify-and-regenerate* mechanism (bottom bar) enforces strict validity via iterative self-correction at every stage to prevent error propagation down the causal chain.

3 The Patient-Zero Framework

We propose **Patient-Zero**, a framework designed to synthesize realistic patient agents without referencing real-world medical records. Formally, we define the framework through two coupled modules: a *Medically-Aligned Hierarchical Synthesis* protocol (Section 3.1) that constructs static patient records, and an *Dual-Track Cognitive Memory System* (Section 3.2) that controls the agent behavior and memory evolution via logic verification.

3.1 Medically-Aligned Hierarchical Synthesis

Traditional approaches formulate patient generation as a derivative task, using existing real-world patient records $p(x_{patient}|x_{record})$. In contrast, **Patient-Zero** models the generation as an *ab initio* synthesis process grounded in abstract medical concepts. Let $\mathcal{K}$ denote a medical knowledge base. The synthesis of a patient record $\mathcal{P}$ is factorized into a chain of conditional dependencies:

$$p(\mathcal{P}|c) = p(\mathcal{E}|\mathcal{S},\mathcal{B},\mathcal{O}) \cdot p(\mathcal{S}|\mathcal{B},\mathcal{O}) \cdot p(\mathcal{B}|\mathcal{O}) \cdot p(\mathcal{O}|c) \quad (1)$$

where $c \in \mathcal{K}$ represents the abstract disease concept (base knowledge seed). The process unfolds in four stages, as illustrated in Figure 2. Generated context in every stages undergo a *verify-and-regenerate* process, which is validated by LLM against a set of criteria, and would be discarded and regenerated if it fails to meet clinical standards.

Stage I: Disease Outline Standardization ($c \to \mathcal{O}$) Given a disease concept c and its associated unstructured medical data retrieved from authoritative sources[1], we utilize an LLM to reconstruct a standardized disease outline $\mathcal{O}$. This stage functions as a *knowledge standardization filter* which parses noisy input data into a unified structured schema anchored in clinical guidelines, serving as a reliable reference for subsequent generation.

Stage II: Epidemiological Attribute Permutation ($\mathcal{O} \to \mathcal{B}$) To ensure the synthesized population strictly aligns with real-world epidemiological distributions, we employ *Stratified Sampling with Constraints* to explicitly decouple attribute selection from textual generation. Specifically, we organize the high-dimensional attribute space $\mathcal{A}$ into a hierarchical taxonomy comprising four core dimensions based on the Social Determinants of Health (SDOH) (World Health Organization, 2024). Each dimension $\mathcal{A}_k$ contains a set of specific fine-grained variables $\{x_1, x_2, \ldots, x_m\}$ designed to comprehensively capture components: **1) Biological and Demographic ($\mathcal{A}_{bio}$):** Age strata, biological sex, physiological status, and ethnicity. **2) Socioeconomic ($\mathcal{A}_{soc}$):** Geographic setting, and socioeconomic status indicators. **3) Behavioral and Lifestyle ($\mathcal{A}_{beh}$):** Substance use, communication, dietary and activity patterns.

Formally, we construct the patient vector $\mathbf{v}$ by concatenating the attribute values sampled from each dimension: $\mathbf{v} = \mathbf{v}_{bio} \oplus \mathbf{v}_{soc} \oplus \mathbf{v}_{fam} \oplus \mathbf{v}_{beh}$, where each sub-vector $\mathbf{v}_k$ consists of sampled values for the specific variables in that category. Let $\mathcal{D}_{prior}$ denote the set of marginal distributions for these sub-attributes derived from authoritative sources[1]. The sampling is formulated as:

$$\mathbf{v} \sim \prod_k P(\mathbf{v}_k|\mathcal{O},\mathcal{D}_{prior}) \cdot \mathbb{I}(\textsc{Valid}(\mathbf{v})) \quad (2)$$

where $P(\mathbf{v}_k|\mathcal{O},\mathcal{D}_{prior})$ denotes the set of categorical distributions weighted by real-world prevalence,

[1]Data sources include verified medical literature, reports and encyclopedias (e.g., Wikipedia Medical Portal).

and $\mathbb{I}(\text{VALID}(\mathbf{v}))$ is an indicator function that performs rejection sampling to eliminate logical inconsistencies (e.g., *Male* $\wedge$ *Pregnant*). Conditioned on this fine-grained vector $\mathbf{v}$, the textual patient profile $\mathcal{B}$ is generated. This stage ensures **Patient-Zero** covers the *"long tail"* of diverse populations while adhering to clinical guidelines. Detailed attribute taxonomies are provided in Appendix B.1.

Stage III: Symptom Progression $(\mathcal{B}, \mathcal{O} \rightarrow \mathcal{S})$ To align with the standard clinical workflow, this stage generates the patient-specific symptoms $\mathcal{S}$ conditioned on the patient profile $\mathcal{B}$ and disease constraints $\mathcal{O}$. We model symptoms as *dynamic trajectories* characterized by onset, duration, severity, and triggers. Symptom generation is conditioned on patient-specific attributes, creating a logical causal link between the patient's background and their clinical presentation.

Stage IV: Granular Clinical Expansion $(\mathcal{S}, \mathcal{B}, \mathcal{O} \rightarrow \mathcal{E})$ The final stage generates detailed clinical examination results $\mathcal{E}$ based on the specific symptoms $\mathcal{S}$ observed in the previous stage, the patient profile $\mathcal{B}$, and the clinical criteria defined in $\mathcal{O}$. We enforce quantitative fidelity by instructing the model to generate specific metrics rather than vague descriptions. The final synthetic record is defined as $\mathcal{P} = \{\mathcal{B}, \mathcal{S}, \mathcal{E}\}$, where $\mathcal{O}$ serves as the latent structural scaffold. The complete synthetic patient record is provided in Appendix F.

3.2 Dual-Track Cognitive Memory System

Bridging the gap between static patient records and dynamic clinical encounters requires transforming the record $\mathcal{P}$ into an embodied agent capable of coherent dialogue. A central challenge in this transformation is the ***Stability-Plasticity Dilemma***[2] (Mermillod et al., 2013): the agent must strictly adhere to the clinical guidelines and memory facts (*stability*) while naturally interacting for unscripted scenarios (*plasticity*). To resolve this, we propose a Dual-Track Cognitive Memory System (Figure 3). Drawing upon the cognitive framework of *declarative memory* (Riedel and Blokland, 2015), we operationalize agent memory not as a monolithic buffer, but as a processing *continuum* between two distinct

[2]Originating in cognitive science and neural networks, the ***Stability-Plasticity Dilemma*** refers to the computational trade-off between the ability of a system to retain existing knowledge (*stability*) and its capacity to acquire new information without overriding valid prior memories (*plasticity*). In our context, this manifests as the tension between rigid medical adherence and open-ended conversational adaptability.

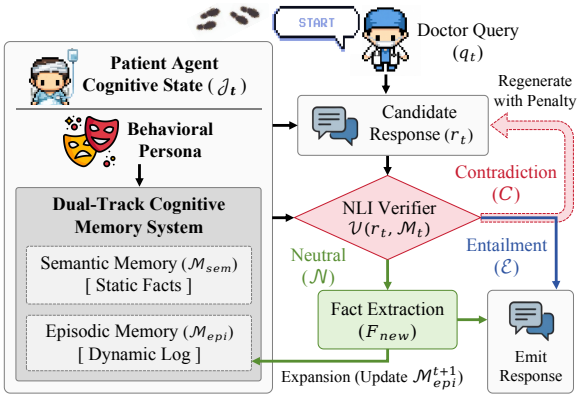


Figure 3: **The Dual-Track Cognitive Memory System.** This module integrates static semantic memory ($\mathcal{M}_{sem}$) and dynamic episodic memory ($\mathcal{M}_{epi}$) to drive coherent dialogue. The *NLI-Verifier* acts as a logical gatekeeper, evaluating candidate responses (r_t) against atomic memory ($\mathcal{M}_t$). By regenerating on Contradictions ($\mathcal{C}$), preserving state on Entailments ($\mathcal{E}$), and expanding on Neutral ($\mathcal{N}$) information, this closed-loop effectively resolves the ***Stability-Plasticity Dilemma***.

levels of abstraction: *semantic memory* for *stability* and *episodic memory* for *plasticity* (Squire, 2004; Tulving, 2002; Hu et al., 2025).

1) Semantic Memory ($\mathcal{M}_{sem}$): In neuroscience, *semantic memory* retains general factual knowledge independent of the specific occasion of acquisition (Squire, 2004). In our framework, we instantiate $\mathcal{M}_{sem}$ using the generated patient record $\mathcal{P}$, serving as the ground-truth anchor to ensure consistency over time. To enable granular logical reasoning, we decompose the complex narrative record $\mathcal{P}$ into a discrete set of atomic propositions $\mathcal{M}_{sem} = \{f_1, f_2, \ldots, f_n\}$, preventing "lost-in-the-middle" phenomena during logical verification.

2) Episodic Memory ($\mathcal{M}_{epi}$): Complementing the semantic layer, *episodic memory* stores personally experienced events associated with specific temporal contexts (Tulving, 2002). We model $\mathcal{M}_{epi}$ as a dynamic log of the ongoing clinical dialogue. This track supports consistency by allowing the agent to mentally re-experience past events, and progressively aligning responses with the doctor's inquiry logic. Unlike the static $\mathcal{M}_{sem}$, this track captures the interaction histories as episodic traces (Zhong et al., 2024; Chhikara et al., 2025).

3.2.1 NLI-Driven Interactive Simulation

To transform these memory structures into dynamic behavior, we define the agent cognitive state at turn t as $\mathcal{J}_t = (\mathcal{M}_{sem} \cup \mathcal{M}_{epi}^t, \Psi)$. Here, Ψ represents a behavioral persona (e.g., *Reserved, Verbose, Up-*

set) injected following the taxonomy in Wang et al. (2024), directly mapping to real communication barriers (Lizée et al., 2024). As highlighted in recent surveys, a critical challenge is preventing hallucinations during the transformation of raw events into memory facts (Tan et al., 2025b,a). We address this via a *NLI-Verifier* that enforces *stability* by rejecting hallucinations, and enables *plasticity* by integrating new non-conflicting details.

Persona and Logical Consistency At interaction turn t, let r_t denote the candidate response generated by the patient agent based on the cognitive state $\mathcal{J}_t$ and the doctor's query q_t. We introduce an *NLI-Verifier* $\mathcal{V}(r_t, \mathcal{M}_t, \Psi)$ that assesses the alignment of r_t against two dimensions: *logical factuality* and *persona adherence*. **1) Logical Factuality:** We define the relation between r_t and a single fact f_i using a discrete logical manifold $\mathcal{L} = \{\textsc{Entail}(\mathcal{E}), \textsc{Contradict}(\mathcal{C}), \textsc{Neutral}(\mathcal{N})\}$:

$$\phi(r_t, f_i) = \begin{cases} \mathcal{E} & \text{if } r_t \models f_i \\ \mathcal{C} & \text{if } r_t \models \neg f_i \\ \mathcal{N} & \text{otherwise} \end{cases} \tag{3}$$

where $\models$ denotes logical entailment. The global verification $\mathbb{V}(r_t, \mathcal{M}_t)$ aggregates these pairwise judgments via a strict priority logic ($\mathcal{C} \succ \mathcal{E} \succ \mathcal{N}$):

$$\mathbb{V}(r_t, \mathcal{M}_t) = \begin{cases} \mathcal{C} & \exists f \in \mathcal{M}_t,\ \phi(r_t, f) = \mathcal{C} \\ \mathcal{E} & [\nexists f \in \mathcal{M}_t,\ \phi(r_t, f) = \mathcal{C}] \\ & \wedge\ [\exists f \in \mathcal{M}_t,\ \phi(r_t, f) = \mathcal{E}] \\ \mathcal{N} & \forall f \in \mathcal{M}_t,\ \phi(r_t, f) = \mathcal{N} \end{cases} \tag{4}$$

where $\mathcal{M}_t = \mathcal{M}_{sem} \cup \mathcal{M}^t_{epi}$. This formulation enforces a strict constraint where any detected contradiction with either semantic facts or prior episodic history leads to a response rejection. **2) Persona Adherence:** To ensure the agent adheres to the behavioral persona Ψ, we treat persona traits as high-level semantic constraints. The *NLI-Verifier* checks if the stylistic attributes of r_t contradict Ψ. Responses that are factually correct but stylistically dissonant are flagged as *persona violation*.

Memory Evolution Policy The agent memory evolves as a deterministic function of the verification outcome. We define the memory transition operator $\mathcal{M}_{t+1} \leftarrow \textsc{Update}(\mathcal{M}_t, r_t)$ conditioned on the verification result $\mathbb{V}(r_t, \mathcal{M}_t)$ as follows:

- **Rejection** ($\mathbb{V} \to \mathcal{C}$; or *persona violation*)**:** The response violates factual constraints in $\mathcal{M}_t$ or deviates from persona Ψ (*violates stability constraint*). It is rejected and regenerated with explicit corrective instructions based on the detected contradiction or persona mismatch.
- **Invariance** ($\mathbb{V} \to \mathcal{E}$)**:** The response r_t is fully grounded in existing memory, and is emitted without modifying the memory state.
- **Expansion** ($\mathbb{V} \to \mathcal{N}$)**:** The response updates new information that is logically consistent with $\mathcal{M}_t$ (*representing plasticity*). Let $\mathcal{F}_{\text{new}}$ denote the set of newly extracted atomic facts from r_t. The *episodic memory* is updated via monotonic expansion: $\mathcal{M}^{t+1}_{epi} \leftarrow \mathcal{M}^t_{epi} \cup \mathcal{F}_{\text{new}}$.

4 Experimental Evaluation

Our evaluation are structured around three core research questions (RQs): **(RQ1) Data Quality:** Does **Patient-Zero** synthesize records that are clinically consistent, diverse, and aligned with real-world distributions? **(RQ2) Interaction Fidelity:** Does our *Dual-Track Cognitive Memory* resolve the *Stability-Plasticity Dilemma*[2] in clinical dialogues? **(RQ3) Downstream Utility:** Does training on our synthetic data improve downstream performance over real-world data constraints?

4.1 Experimental Setup

Datasets We constructed a large-scale synthetic dataset comprising 60,000 patient records spanning six major clinical specialties: Cardiology, Gastroenterology, General Surgery, Neurology, Psychiatry, and Pulmonology. The dataset covers **98 distinct disease types**, ranging from common chronic conditions to critical acute emergencies. The distribution adheres to real-world epidemiological prevalence. Detailed disease and specialty statistics are provided in Appendix B.

Evaluation Metrics A multi-facet protocol is used to assess the three research questions across distinct state-of-the-art baselines: **1) Data Quality:** We evaluate the synthesized records across *Linguistic Quality* (Perplexity, Distinct-4), *Semantic Diversity* (Self-BLEU, Entity Diversity), and *Clinical Validity* (Consistency, Completeness). **2) Interaction Fidelity:** We assess the agent interactive performance via *Factual Fidelity* (Logical Consistency, Factual Recall), *Behavioral Fidelity* (Persona Alignment, Stylistic Stability), *Safety and Robustness* (Hallucination Rate, Inducibility Resistance). **3) Downstream Utility:** Accuracy improvements on MedQA and MMLU. (see Appendix C.1 for baseline configurations and Appendix C.2 for metrics calculation methodologies).

Table 1: Overall performance of Data Quality. Small **gain** subscripts denote the performance gain compared to the ablation baseline using the same backbone, while **real-world data** serves as the Gold Standard.

Method	Backbone	Linguistic Quality		Semantic Diversity		Clinical Validity	
		PPL (↓)	Distinct-4 (↑)	Self-BLEU (↓)	Entity Diversity (↑)	Consis-tency (↑)	Complete-ness (↑)
Real-world Data Baselines							
MIMIC-IV (Johnson et al., 2023)	–	1.50	10.99	85.89	4.15	92.73	45.44
SCRIPT X2B8 (Fenske et al., 2025)	–	2.38	51.06	74.75	0.78	96.12	56.56
CMA Base (CMA)	–	4.82	83.63	42.20	27.02	96.78	91.35
Synthetic Data Baselines							
Synthea (Walonoski et al., 2017)	Rule-based	12.20	20.00	93.24	5.84	79.66	35.32
LDP-GAN (Gwon et al., 2024)	GAN	10.86	30.70	91.18	1.10	85.89	46.40
Avatar (Guillaudeux et al., 2023)	FAMD + KNN	9.03	48.43	70.42	3.74	95.22	42.84
MERA-Mistral (Ibrahim et al., 2025)	Mistral-7B-v0.3	3.88	50.43	77.61	6.18	75.84	69.20
MERA-Llama (Ibrahim et al., 2025)	Llama-3-70B	3.51	50.57	77.09	8.27	82.15	97.00
MERA-Qwen (Ibrahim et al., 2025)	Qwen-2.5-32B	**3.40**	66.02	68.45	7.88	88.88	95.80
Ablation Baselines							
Direct (w/o Hierarchical Synthesis)	GPT-5	9.19	61.83	59.16	21.69	70.76	75.75
	Gemini-2.5-Pro	7.53	62.00	70.61	19.98	79.42	76.13
	Claude-Sonnet-4	9.16	65.89	61.83	21.78	71.35	61.50
Patient-Zero	GPT-5	4.48 $_{-4.71}$	**76.51** $_{+14.68}$	**52.45** $_{-6.71}$	**23.72** $_{+2.03}$	**99.12** $_{+28.36}$	**99.88** $_{+24.13}$
	Gemini-2.5-Pro	4.23 $_{-3.30}$	74.15 $_{+12.15}$	57.64 $_{-12.97}$	23.36 $_{+3.38}$	98.89 $_{+19.47}$	98.91 $_{+22.78}$
	Claude-Sonnet-4	4.37 $_{-4.79}$	71.35 $_{+5.46}$	55.67 $_{-6.16}$	22.55 $_{+0.77}$	98.23 $_{+26.88}$	98.47 $_{+36.97}$

4.2 (RQ1) Quality of Synthetic Data

Patient-Zero outperforms the real-world data approximations and synthetic baselines in Table 1, consistently yielding substantial improvements in clinical validity (>98%), demonstrating the generalizability across different backbone architectures.

1) Breaking the Privacy-Utility Trade-off A critical challenge in patient data synthesis is the trade-off between privacy protection and data utility. Traditional synthetic baselines often suffer from mode collapse and clinical accuracy degradation due to privacy noise injection. In contrast, **Patient-Zero** effectively resolves this dilemma. By decoupling privacy from generation via *ab initio* synthesis, we achieve near-perfect clinical validity (>98%), while maintaining superior diversity.

2) Superiority over Direct Synthesis To validate the necessity of our *Hierarchical Synthesis* framework, we compare **Patient-Zero** against its ablation baseline. Direct synthesis often yields outputs with high repetition and low clinical validity, as indicated by the **gain** subscripts in Table 1. This confirms that our hierarchical constraints effectively mitigate textual degeneration, enable models to generate non-stereotypical patient narratives.

3) Holistic Metric Balance While MERA baselines achieve lower Perpelxity (PPL) at the cost of lexical richness, with repetitive generic medical phrasing. In contrast, **Patient-Zero** achieves significantly higher lexical diversity (+15.45 in Distinct-4 over MERA-Qwen), while does not degrade clinical validity, confirming that our data represent ***rich, non-stereotypical clinical narratives rather than merely optimizing for syntactic predictability.***

4.3 (RQ2) Fidelity of Interaction

We operationalize Interaction Fidelity as the resolution of the *Stability-Plasticity Dilemma*. We simulate multi-turn diagnostic sessions using the advanced reasoning capabilities of Doctor-R1 (Lai et al., 2025), explicitly instructed to execute adversarial probing to stress-test the patient agents.

1) Resolving the Stability-Plasticity Dilemma Our framework effectively resolves the trade-off between clinical consistency and conversational fidelity (Table 2), surpassing state-of-the-art agent baselines across all metrics. The substantial gain of +18.64% and +25.39% in consistency and recall in the ablation confirms that our *Dual-Track Cognitive Memory* is essential for long-horizon coherence.

2) Mitigating Compliance Bias A critical vulnerability in medical simulation is *compliance bias*, where patient agents hallucinate symptoms solely to align with the doctor suggestions. Our framework outperforms EvoPatient (Du et al., 2024) and Patient-Ψ (Wang et al., 2024) in both Hallucination Rate and Inducibility Resistance, while improving +28.05% resistance and reduce 8.63% hallucination over the ablation. This validates the efficacy of our *Dual-Track Cognitive Memory* by proactively rejecting hallucinated responses.

Table 2: Overall performance of Interaction Fidelity. Small **gain** subscripts denote the absolute performance gain compared to the ablation baseline (without *Dual-Track Cognitive Memory*).

Method	Factual Fidelity (Cognitive)		Behavioral Fidelity (Persona)		Safety & Robustness	
	Logical Consistency (↑)	Factual Recall (↑)	Persona Alignment (↑)	Stylistic Stability (↑)	Hallucination Rate (↓)	Inducibility Resistance (↑)
State-of-the-art Medical Agents						
EvoPatient (Du et al., 2024)	87.49	9.78	80.50	32.95	5.01	94.87
AI Hospital (Fan et al., 2024)	97.95	16.68	78.80	29.33	6.67	86.36
MediQ (Li et al., 2024)	81.83	27.04	77.20	28.03	8.46	92.86
Patient-Ψ (Wang et al., 2024)	78.75	10.10	71.20	40.74	2.00	88.89
Architectural Ablations						
Unstructured Narrative	90.00	15.39	60.20	61.11	4.55	87.72
Static Structured Schema	93.18	14.41	67.20	64.60	2.73	79.90
without *Dual-Track Cog. Mem.*	81.36	7.88	61.10	46.61	10.45	69.17
Patient-Zero	**100.00** $_{+18.64}$	**33.27** $_{+25.39}$	**83.70** $_{+22.60}$	**75.50** $_{+28.89}$	**1.82** $_{-8.63}$	**97.22** $_{+28.05}$

Table 3: Downstream Performance and Results.

Model	MedQA	MMLU
Med42-v2-8B	62.50	59.50
UltraMedical-8B	74.00	68.50
Baichuan-M2-32B	82.50	84.00
Qwen3-8B (Base Model)	61.00	69.50
Qwen3-8B + Real Data	83.50	**85.50**
Qwen3-8B + **Patient-Zero**	**85.00**	83.00

4.4 (RQ3) Downstream Clinical Utility

To evaluate downstream utility, we trained Qwen3-8B (Yang et al., 2025) on **Patient-Zero** and assessed it on two clinical reasoning benchmarks: MedQA (Jin et al., 2020) and MMLU (Hendrycks et al., 2021). As shown in Table 3, training on our synthetic data yields substantial gains over the base model (+24.0% on MedQA; +14.5% on MMLU). Under the same setup, a model trained on real data performs similarly, while model trained on **Patient-Zero** surpasses medical-LLM baselines, indicating that our synthetic data provides training utility comparable to real data (see Appendix C.3 for detailed downstream training implementation).

4.5 Human Expert Evaluation

To assess the clinical validity of our synthetic data, we employed senior licensed physicians from top-tier hospitals to evaluate **Patient-Zero** against real clinical data and baseline models. We conclude a key finding from their feedback: ***real-world data often deviates from the "ideal" clinical standard***, exhibiting issues such as missing fields, chaotic logic, or structural noise. In contrast, our framework generates idealized clinical narratives with superior structural integrity. This section reports the results of the blind discrimination test and the multi-dimensional quality ratings. Detailed annotation protocols, qualitative case studies, and statistical breakdowns are provided in Appendix E.

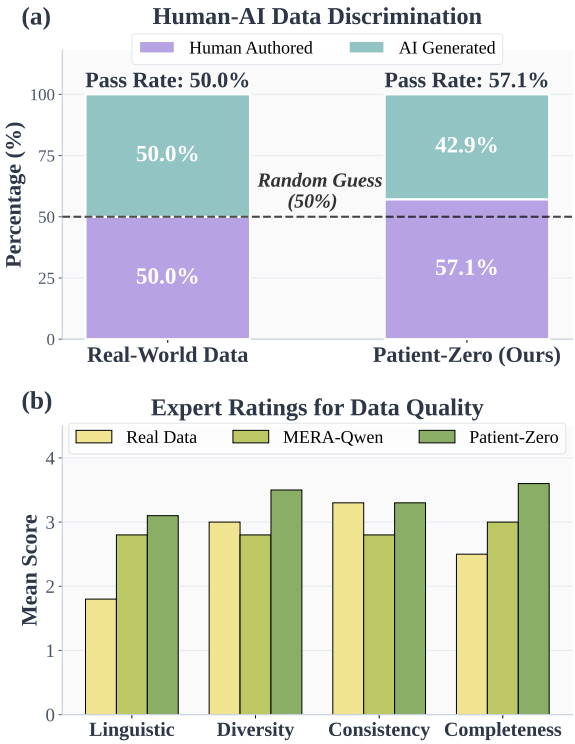


Figure 4: **Expert Evaluation Results.** (a) Senior licensed physicians showed near-chance discrimination between real and synthetic records, with **Patient-Zero** judged more frequently as human-authored. (b) Experts rated our framework highest in overall clinical quality.

Human-AI Data Discrimination To assess indistinguishability, we conducted a blinded discrimination test. As shown in Figure 4(a), **Patient-Zero** was classified as human-authored in 57.1% of cases, surpassing the actual real data (50.0%). This result reveals a critical insight: ***physicians rely on a quality heuristic to distinguish human-authored content from AI generation.*** Experts tend to classify data with high completeness and logical flow as human-authored, while flagging records with noise, incompleteness, and chaotic formatting as AI-generated. Consequently, 50% of real human-authored records were misclassified as AI due to

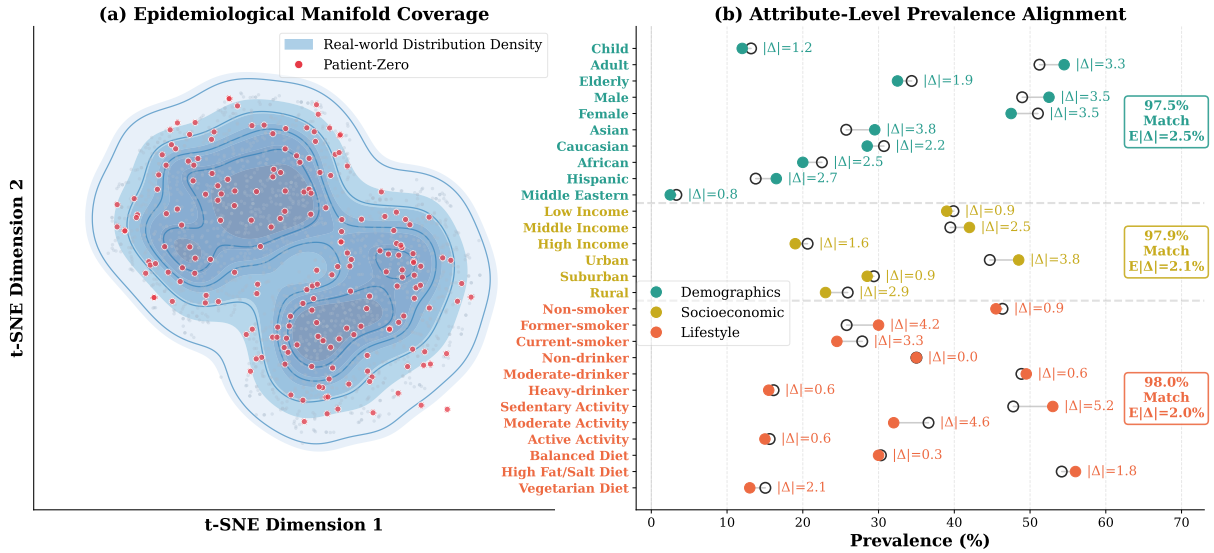


Figure 5: **Holistic Epidemiological and Semantic Alignment. (a) Epidemiological Manifold Coverage:** The tight overlap of t-SNE visualization in the high-dimensional semantic space indicates that our synthetic data captures the comprehensive epidemiological manifold without mode collapse. **(b) Attribute-Level Prevalence Alignment:** Prevalence comparison across three dimensions. Hollow circles (○) represent real-world baseline; solid circles (●) represent **Patient-Zero**; the connecting lines indicate alignment gaps. Minimal absolute differences demonstrate that Patient-Zero faithfully reconstructs complex real-world demographic and behavioral profiles.

their inherent flaws. **Patient-Zero** shows alignment with the physicians' stereotype of a *good human-authored patient record*, demonstrating ***no significant perceptible difference remains between our synthetic data and real clinical records***.

Expert Quality Ratings Licensed physicians evaluate data quality across four dimensions on a Likert scale. **Patient-Zero** achieves the highest overall score, outperforming both the MERA-Qwen (Ibrahim et al., 2025) synthetic baseline and real data (CMA) in Figure 4(b). These results strongly validate the findings in Table 1, showing that our method delivers higher clinical quality visible to human experts. Statistical analysis confirms that **Patient-Zero** achieves significant gains over real data in Linguistic Quality and Completeness (*Mann-Whitney U*, $p < 0.05$). This conclusion is reliable given the high inter-rater consistency in our rater bias analysis (*Kruskal-Wallis*, $p = 0.89$).

5 Distribution Analysis

Holistic Epidemiological Alignment A fundamental imperative in synthetic data generation is maximizing the *coverage* of the underlying data manifold while minimizing divergence from the ground-truth distribution. We rigorously quantify the congruence between the synthetic distribution $\hat{\mathcal{P}}$ and the real-world prior $\mathcal{P}^*$ from both geometric and statistical perspectives. **1) Epidemiological Manifold Coverage:** As visualized in the t-SNE embedding (Figure 5a), **Patient-Zero** demonstrates comprehensive coverage of the high-dimensional topological support, effectively approximating the continuous density of the real-world manifold. **2) Attribute-Level Prevalence:** Figure 5(b) quantifies the deviation between the empirical synthetic probability mass functions and the ground-truth baselines. We validate this alignment via hypothesis testing (detailed in Appendix D). Chi-square tests across all marginal distributions yield no statistically significant deviations ($p > 0.05$), with key covariates like ethnicity ($p = 0.49$) and age ($p = 0.66$) showing high conformity. Applying *Fisher's combined probability test* to the *global null hypothesis* $H_0 : \hat{\mathcal{P}} = \mathcal{P}^*$ yields an aggregate p-value of 0.86, indicating that ***the synthetic cohort is statistically indistinguishable from the real-world population***, validating high-fidelity reconstruction without privacy leakage.

6 Conclusion

This paper presents **Patient-Zero**, an *ab initio* framework designed to bridge the gap between static medical record generation and dynamic patient simulation. By integrating *Medically-Aligned Hierarchical Synthesis* with a *Dual-Track Cognitive Memory*, our approach resolves the privacy–utility paradox while remaining aligned with real-world epidemiological distributions. Empirical results demonstrate that **Patient-Zero** estab-

lishes a new standard for clinical fidelity, generating synthetic data that experts perceive as authentic human-authored narratives.

Limitations

Methodology-wise, the primary limitation of our work is the *idealization bias* inherent in our guideline-anchored synthesis. While this structural integrity is favored by physicians, empirical studies on how this perfect data affects model robustness against real-world clinical noise shall be valuable. Scope-wise, we have not systematically extended the framework to *multimodal* clinical data or validated the system in prospective clinical trials due to ethical and safety constraints. In future works, we plan to focus on constructing *holistic virtual patients* that integrate visual and signal-based modalities. Additionally, we aim to theoretically uncover the relationship between synthetic data scale and downstream reasoning performance, specifically determining the optimal ratio of synthetic-to-real data required to maximize clinical utility without incurring hallucination risks.

Ethics Statement

Patient-Zero generates data *ab initio* based on abstract medical knowledge, hence no real patient data or private health information (PHI) is utilized or exposed in the development of this framework. This work strictly adheres to data privacy standards and poses no risk of re-identification. However, our approach is a research prototype intended for educational simulation and model training; the generated records are synthetic and should not be used as a substitute for real clinical data in healthcare decision-making without further validation.

Human Evaluation Protocol We conducted a human evaluation involving two senior licensed physicians from top-tier hospitals. These experts were tasked with a blinded discrimination test to distinguish between synthetic and real-world records, and also a multi-dimensional quality assessment. The evaluation was strictly double-blind. The annotators are unaware of the data sources and were instructed to evaluate based solely on defined qualitiative dimensions. We explicitly state that all participating expert annotators were fully informed about the study's purpose and data usage. They provided explicit consent via the user agreement on the annotation interface before commencing the evaluation tasks.

Annotator Welfare and Consent The participating physicians were fully informed about the purpose of the study and the tasks involved. All participation was voluntary, and the experts were compensated at a professional rate commensurate with their experience and local labor standards. No personal information regarding the evaluators was retained beyond their professional credentials required for the study.

Potential Risks We acknowledge potential risks inherent in synthetic medical data, including the possibility of hallucinated medical details or the amplification of biases present in the seed knowledge bases (e.g., Wikipedia). While our distribution analysis indicates high alignment with real-world statistics, synthetic cohorts may not fully capture rare anomalies or complex comorbidities found in organic populations. We release this dataset to facilitate privacy-safe research while emphasizing that models trained on this data should undergo safety testing before any deployment.

Data Governance and Compliance This study adheres to strict data governance protocols. We utilized the MIMIC-IV dataset (Johnson et al., 2023) solely for benchmarking and distribution comparison purposes. Access to MIMIC-IV was obtained under the *PhysioNet Credentialed Health Data Use Agreement* (DUA). All authors accessing this data have completed the required CITI Program training in "Data or Specimens Only Research". For open-source datasets and models, we strictly complied with their respective licenses and utilized them exclusively for academic research. Crucially, our proposed framework, **Patient-Zero**, generates synthetic records *ab initio* from abstract medical guidelines. Therefore, the released synthetic corpus contains no Personally Identifiable Information (PII) or Protected Health Information (PHI) derived from real patients, eliminating the risk of privacy leakage or re-identification.

Reproducibility Statement

To facilitate future research and ensure reproducibility, we will publicly release the complete **Patient-Zero** framework code, the generated synthetic dataset, and the evaluation scripts on GitHub and Hugging Face. Our generation pipeline utilizes accessible Large Language Models and public medical knowledge bases. All downstream evaluations were performed on established public benchmarks,

including MedQA and MMLU. A detailed breakdown of the hierarchical generation prompts, and experimental hyperparameters is provided in Appendix C.

References

Chinese medical ace base (yiigle). Yiigle Chinese Medical Database. Accessed: 2025.

F. Chen, Y. Li, Y. Chen, Z. Bian, L. Duo, Q. Zhou, L. Zhang, and ADVANCED Working Group. 2025. Strategies for the analysis and elimination of hallucinations in artificial intelligence generated medical knowledge. *Journal of Evidence-Based Medicine*, 18(3):e70075.

Yan Chen and Pouyan Esmaeilzadeh. 2024. Generative ai in medical practice: In-depth exploration of privacy and security challenges. *Journal of Medical Internet Research*, 26:e53008.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. 2025. Mem0: Building production-ready ai agents with scalable long-term memory. *Preprint*, arXiv:2504.19413.

Clément Christophe, Praveen K Kanithi, Tathagata Raha, Shadab Khan, and Marco AF Pimentel. 2024. Med42-v2: A suite of clinical llms.

Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2nd edition. Lawrence Erlbaum Associates, Hillsdale, NJ.

David A. Cook, Jesper Overgaard, V. Shane Pankratz, Guilherme Del Fiol, and Christopher A. Aakre. 2025. Virtual patients using large language models: Scalable, contextualized simulation of clinician-patient dialogue with feedback. *Journal of Medical Internet Research*, 27:e68486.

Trisha Das, Zifeng Wang, Afrah Shafquat, Mandis Beigi, Jason Mezey, and Jimeng Sun. 2024. Synrl: Aligning synthetic clinical trial data with human-preferred clinical endpoints using reinforcement learning. *Preprint*, arXiv:2411.07317.

Zhuoyun Du, Lujie Zheng, Renjun Hu, Yuyang Xu, Xiawei Li, Ying Sun, Wei Chen, Jian Wu, Haolei Cai, and Haohao Ying. 2024. Llms can simulate standardized patients via agent coevolution. *Preprint*, arXiv:2412.11716.

Zhihao Fan, Jialong Tang, Wei Chen, Siyuan Wang, Zhongyu Wei, Jun Xi, Fei Huang, and Jingren Zhou. 2024. Ai hospital: Benchmarking large language models in a multi-agent medical interaction simulator. *Preprint*, arXiv:2402.09742.

Yinqiu Feng, Bo Zhang, Lingxi Xiao, Yutian Yang, Tana Gegen, and Zexi Chen. 2024. Enhancing medical imaging with gans synthesizing realistic images from limited data. *Preprint*, arXiv:2406.18547.

Samuel W. Fenske, Alec Peltekian, Mengjia Kang, Nikolay S. Markov, Mengou Zhu, Kevin Grudzinski, Melissa J. Bak, Anna Pawlowski, Vishu Gupta, Yuwei Mao, Stanislav Bratchikov, Thomas Stoeger, Luke V. Rasmussen, Alok N. Choudhary, Alexander V. Misharin, Benjamin D. Singer, G. R. Scott Budinger, Richard G. Wunderink, Ankit Agrawal, and 2 others. 2025. Developing and validating machine learning models to predict next-day extubation. *Scientific Reports*, 15(1):27552.

Mauro Giuffrè and Dennis L. Shung. 2023. Harnessing the power of synthetic data in healthcare: Innovation, application, and privacy. *npj Digital Medicine*, 6(1):186.

Andre Goncalves, Priyadip Ray, Braden Soper, Jennifer Stevens, Linda Coyle, and Ana Paula Sales. 2020a. Generation and evaluation of synthetic patient data. *BMC Medical Research Methodology*, 20(1):108.

Andre Goncalves, Priyadip Ray, Braden Soper, Jennifer Stevens, Linda Coyle, and Ana Paula Sales. 2020b. Generation and evaluation of synthetic patient data. *BMC Med. Res. Methodol.*, 20(1).

Leonhard Graf, Philipp Sykownik, Gerhard Gradl-Dietsch, and Maic Masuch. 2024. Towards believable and educational conversations with virtual patients. *Frontiers in Virtual Reality*, 5:1377210.

Morgan Guillaudeux, Olivia Rousseau, Julien Petot, Zineb Bennis, Charles-Axel Dein, Thomas Goronflot, Nicolas Vince, Sophie Limou, Matilde Karakachoff, Matthieu Wargny, and Pierre-Antoine Gourraud. 2023. Patient-centric synthetic data generation, no reason to risk re-identification in biomedical data analysis. *npj Digital Medicine*, 6(1):37.

Xu Guo and Yiqiang Chen. 2024. Generative ai for synthetic data generation: Methods, challenges and the future. *Preprint*, arXiv:2403.04190.

Hansle Gwon, Imjin Ahn, Yunha Kim, Hee Jun Kang, Hyeram Seo, Heejung Choi, Ha Na Cho, Minkyoung Kim, JiYe Han, Gaeun Kee, Seohyun Park, Kye Hwa Lee, Tae Joon Jun, and Young-Hak Kim. 2024. Ldp-gan : Generative adversarial networks with local differential privacy for patient medical records synthesis. *Computers in Biology and Medicine*, 168:107738.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Yuyang Hu, Shichun Liu, Yanwei Yue, Guibin Zhang, Boyang Liu, Fangyi Zhu, Jiahang Lin, Honglin Guo, Shihan Dou, Zhiheng Xi, Senjie Jin, Jiejun Tan, Yanbin Yin, Jiongnan Liu, Zeyu Zhang, Zhongxiang Sun, Yutao Zhu, Hao Sun, Boci Peng, and 28 others. 2025. Memory in the age of ai agents. *Preprint*, arXiv:2512.13564.

Ahmed Ibrahim, Abdullah Khalili, Maryam Arabi, Aamenah Sattar, Abdullah Hosseini, and Ahmed Serag.

2025. Mera: Medical electronic records assistant. *Machine Learning and Knowledge Extraction*, 7(3).

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Preprint*, arXiv:2009.13081.

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

Fabian Sven Karst, Sook-Yee Chong, Abigail A. Antenor, Enyu Lin, Mahei Manhai Li, and Jan Marco Leimeister. 2024. Generative ai for banks: Benchmarks and algorithms for synthetic financial transaction data. *Preprint*, arXiv:2412.14730.

Clemens Scott Kruse, Brenna Smith, Hannah Vanderlinden, and Alexandra Nealand. 2017. Security techniques for the electronic health records. *Journal of Medical Systems*, 41(8).

Yunghwei Lai, Kaiming Liu, Ziyue Wang, Weizhi Ma, and Yang Liu. 2025. Doctor-r1: Mastering clinical inquiry with experiential agentic reinforcement learning. *Preprint*, arXiv:2510.04284.

Lucas Lange, Nils Wenzlitschke, and Erhard Rahm. 2024. Generating synthetic health sensor data for privacy-preserving wearable stress detection. *Sensors*, 24(10).

Jingoo Lee, Kyungho Lim, Young-Chul Jung, and Byung-Hoon Kim. 2025. Psyche: A multi-faceted patient simulation framework for evaluation of psychiatric assessment conversational agents. *Preprint*, arXiv:2501.01594.

Jin Li, Benjamin J. Cairns, Jingsong Li, and Tingting Zhu. 2023a. Generating synthetic mixed-type longitudinal electronic health records for artificial intelligent applications. *npj Digital Medicine*, 6(1):98.

Shuyue Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S. Ilgen, Emma Pierson, Pang Wei Koh, and Yulia Tsvetkov. 2024. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. In *Advances in Neural Information Processing Systems*, volume 37, pages 28858–28888. Curran Associates, Inc.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023b. Synthetic data generation with large language models for text classification: Potential and limitations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10443–10461, Singapore. Association for Computational Linguistics.

Tengfei Liu, Jiapu Wang, Yongli Hu, Mingjie Li, Junfei Yi, Xiaojun Chang, Junbin Gao, and Baocai Yin. 2024. Hc-llm: Historical-constrained large language models for radiology report generation. *Preprint*, arXiv:2412.11070.

Antoine Lizée, Pierre-Auguste Beaucoté, James Whitbeck, Marion Doumeingts, Anaël Beaugnon, and Isabelle Feldhaus. 2024. Conversational medical ai: Ready for practice. *Preprint*, arXiv:2411.12808.

Ryan Louie, Ananjan Nandi, William Fang, Cheng Chang, Emma Brunskill, and Diyi Yang. 2024. Roleplay-doh: Enabling domain-experts to create LLM-simulated patients via eliciting and adhering to principles. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10570–10603, Miami, Florida, USA. Association for Computational Linguistics.

Bruno Macedo, Inês Ribeiro Vaz, and Tiago Taveira Gomes. 2024. MedGAN: optimized generative adversarial network with graph convolutional networks for novel molecule design. *Sci. Rep.*, 14(1):1212.

Martial Mermillod, Aurélia Bugaiska, and Patrick Bonin. 2013. The stability-plasticity dilemma: investigating the continuum from catastrophic forgetting to age-limited learning effects. *Frontiers in Psychology*, Volume 4 - 2013.

Md Mahadi Hasan Nahid and Sadid Bin Hasan. 2024. Safesynthdp: Leveraging large language models for privacy-preserving synthetic data generation using differential privacy. *Preprint*, arXiv:2412.20641.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Wim J. Riedel and Arjan Blokland. 2015. Declarative memory. In *Handbook of Experimental Pharmacology*, volume 228, pages 215–236. Springer.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2025. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, EuroSys '25, page 1279–1297. ACM.

Larry R. Squire. 2004. Memory systems of the brain: A brief history and current perspective. *Neurobiology of Learning and Memory*, 82(3):171–177.

Hongda Sun, Hongzhan Lin, and Rui Yan. 2024. Collaborative synthesis of patient records through multi-visit health state inference. AAAI'24/IAAI'24/EAAI'24. AAAI Press.

Haoran Tan, Zeyu Zhang, Chen Ma, Xu Chen, Quanyu Dai, and Zhenhua Dong. 2025a. MemBench: Towards more comprehensive evaluation on the memory of LLM-based agents. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19336–19352, Vienna, Austria. Association for Computational Linguistics.

Zhen Tan, Jun Yan, I-Hung Hsu, Rujun Han, Zifeng Wang, Long Le, Yiwen Song, Yanfei Chen, Hamid Palangi, George Lee, Anand Rajan Iyer, Tianlong Chen, Huan Liu, Chen-Yu Lee, and Tomas Pfister. 2025b. In prospect and retrospect: Reflective memory management for long-term personalized dialogue agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8416–8439, Vienna, Austria. Association for Computational Linguistics.

M2 Team, Chengfeng Dou, Chong Liu, Fan Yang, Fei Li, Jiyuan Jia, Mingyang Chen, Qiang Ju, Shuai Wang, Shunya Dang, Tianpeng Li, Xiangrong Zeng, Yijie Zhou, Chenzheng Zhu, Da Pan, Fei Deng, Guangwei Ai, Guosheng Dong, Hongda Zhang, and 15 others. 2025. Baichuan-m2: Scaling medical capability with large verifier system. *Preprint*, arXiv:2509.02208.

Margaux Tornqvist, Jean-Daniel Zucker, Tristan Fauvel, Nicolas Lambert, Mathilde Berthelot, and Antoine Movschin. 2024. A text-to-tabular approach to generate synthetic patient data using llms. *Preprint*, arXiv:2412.05153.

Endel Tulving. 2002. Episodic memory: From mind to brain. *Annual Review of Psychology*, 53:1–25.

Jason Walonoski, Mark Kramer, Joseph Nichols, Andre Quina, Chris Moesel, Dylan Hall, Carlton Duffett, Kudakwashe Dube, Thomas Gallagher, and Scott McLachlan. 2017. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*, 25(3):230–238.

Ruiyi Wang, Stephanie Milani, Jamie C. Chiu, Jiayin Zhi, Shaun M. Eack, Travis Labrum, Samuel M Murphy, Nev Jones, Kate V Hardy, Hong Shen, Fei Fang, and Zhiyu Chen. 2024. PATIENT-ψ: Using large language models to simulate patients for training mental health professionals. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12772–12797, Miami, Florida, USA. Association for Computational Linguistics.

Robert Wasenmüller, Kevin Hilbert, and Christoph Benzmüller. 2024. Script-based dialog policy planning for llm-powered conversational agents: A basic architecture for an "ai therapist". *Preprint*, arXiv:2412.15242.

World Health Organization. 2024. Social determinants of health. https://www.who.int/health-topics/social-determinants-of-health. Accessed: 2024-05-20.

Zerui Xu, Fang Wu, Yuanyuan Zhang, and Yue Zhao. 2025. Retrieval-reasoning large language model-based synthetic clinical trial generation. *Preprint*, arXiv:2410.12476.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

H. Yu, J. Zhou, L. Li, and 1 others. 2025. Simulated patient systems powered by large language model-based ai agents offer potential for transforming medical education. *Communications Medicine*.

Kaiyan Zhang, Sihang Zeng, Ermo Hua, Ning Ding, Zhang-Ren Chen, Zhiyuan Ma, Haoxin Li, Ganqu Cui, Biqing Qi, Xuekai Zhu, Xingtai Lv, Hu Jinfang, Zhiyuan Liu, and Bowen Zhou. 2024. Ultramedical: Building specialized generalists in biomedicine.

Xuanliang Zhang, Dingzirui Wang, Longxu Dou, Qingfu Zhu, and Wanxiang Che. 2025. A survey of table reasoning with large language models. *Frontiers of Computer Science*, 19(9):199348.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 19724–19731.

Yunqi Zhu, Wen Tang, Ying Sun, and Xuebing Yang. 2024. The potential of llms in medical education: Generating questions and answers for qualification exams. *Preprint*, arXiv:2410.23769.

A LLM Usage Statement

Throughout the completion of this work, the Large Language Model (LLM) was used solely for the purpose of refining sentences, improving grammatical accuracy and fluency during the manuscript writing process.

B Data Construction Details

B.1 Hierarchical Attribute Taxonomy

To ensure the synthesized population strictly aligns with real-world epidemiological distributions, we employ a stratified sampling strategy based on the taxonomy detailed in Table 4. This taxonomy comprehensively considers dimensions ranging from biological characteristics to the Social Determinants of Health (SDOH) (World Health Organization, 2024).

Dimension	Attributes and Permutations
Biological & Demographic	**Age Strata**: Child, Adult, Elderly **Biological Sex**: Male, Female **Physiological Status**: Pregnant, Non-pregnant, Post-menopausal **Ethnicity**: Asian, Caucasian, African American, Hispanic, Mixed, Other
Socio-economic	**Geography**: Urban (Metropolitan), Rural, Suburban; Specific Region Constraints **Socioeconomic Status**: Education Level, Occupation Type, Income Tier (Low/Middle/High)
Behavioral & Lifestyle	**Substance Use**: Smoking (Never/Former/Current), Alcohol (None/Moderate/Heavy) **Diet & Activity**: Dietary Pattern (Balanced/High-fat/High-salt); Activity Level (Sedentary/Moderate/Active) **Communication Style**: Plain, Upset, Verbose, Reserved, Tangent, Pleasing **Preferences**: Preference for Modern vs. Traditional Medicine

Table 4: **Hierarchical Taxonomy of Patient Attributes.** This structure is used in the *Patient-Zero* generation framework for attribute permutation to construct diverse patient records.

B.2 Disease Taxonomy Coverage

The **Patient-Zero** dataset covers a wide spectrum of pathologies across six specialties, as shown in Table 5. The selection ensures coverage of diverse clinical presentations, including infectious diseases, malignancies, chronic metabolic disorders, and acute trauma.

C Experimental Configuration and Implementation

C.1 Baselines Details

We categorize our baselines into two distinct groups to strictly align with the evaluation tasks reported in Table 1 and Table 2.

Baselines for Data Quality (RQ1) To evaluate the static quality of generated patient records ($\mathcal{P}$), we compare against four categories of baselines:

- **Real-world Data Baselines:** We employ de-identified clinical records from three sources to serve as the topological "Gold Standard" for distribution alignment: **1) MIMIC-IV** (Johnson et al., 2023): Large-scale ICU electronic health records. **2) CMA Base** (CMA): A comprehensive clinical medical association dataset representing general practice distributions.
- **Traditional Synthetic Baselines:** Established non-LLM approaches representing previous generation paradigms: **1) Synthea** (Walonoski et al., 2017): A rule-based engine simulating patient lifespans and medical history. **2) LDP-GAN** (Gwon et al., 2024): A Generative Adversarial Network incorporating Local Differential Privacy. **3) Avatar** (Guillaudeux et al., 2023): A hybrid method combining Factor Analysis of Mixed Data (FAMD) with k-Nearest Neighbors (KNN).
- **LLM-based Synthetic Baselines**: We evaluate the **MERA** (Ibrahim et al., 2025) framework of different open-source backbones to benchmark generation capabilities, including MERA-Mistral (Mistral-7B-v0.3), MERA-Llama (Llama-3-70B), and MERA-Qwen (Qwen-2.5-32B).
- **Ablation Baselines:** To validate the necessity of our *Hierarchical Synthesis*, we compare our framework against an approach without hierarchical constraints, using proprietary models: **GPT-5** (gpt-5-2025-08-07), **Gemini-2.5-Pro**, and **Claude-Sonnet-4** (claude-sonnet-4-20250514).

Baselines for Interaction Fidelity (RQ2) To evaluate the dynamic performance of the patient agent during medical consultations, we compare against state-of-the-art agents and internal architectural ablations:

- **State-of-the-art Medical Agents: 1) EvoPatient** (Du et al., 2024): Uses evolutionary algorithms to iteratively optimize patient profiles for diagnostic realism. **2) AI Hospital** (Fan et al., 2024): A multi-agent framework utilizing role-specific prompts for simulation. **3) MediQ** (Li et al., 2024): A framework emphasizing accurate clinical reasoning and query response. **4) Patient-Ψ** (Wang et al., 2024): A cognitive modeling approach focusing on psychological and behavioral fidelity.
- **Architectural Ablations:** To isolate the contributions of our specific components: **1) Unstructured Narrative:** The agent utilizes raw

Specialty	Covered Diseases
Cardiology	Congenital Heart Disease, Hypertension, Pulmonary Adenocarcinoma, Esophageal Cancer, Fractured Rib, Atherosclerosis, Emphysema, Rheumatic Heart Disease, Acute Myocardial Infarction, Myocardial Ischemia, Heart Failure, Myocarditis, Arrhythmia, Hemothorax, Aortic Stenosis, Mediastinal Emphysema, Hypertensive Crisis, Coronary Artery Disease, Hyperlipidemia
General Surgery	Hyperthyroidism, Mammary Hyperplasia, Cirrhosis, Corn, Hepatocellular Carcinoma, Thyroid Cancer, Colorectal Cancer, Breast Cancer, Gallbladder Polyps, Gallstones, Inguinal Hernia, Acute Appendicitis, Chronic Appendicitis, Pancreatitis, Anal Polyp
Neurology	Stroke, Acute Myelitis, Tic Disorder, Cerebral Arteriosclerosis, Viral Meningitis, Hydrocephalus, Cerebral Ischemia, Leukoencephalopathy, Parkinson's Disease, Trigeminal Neuralgia, Vertigo, Ataxia, Vascular Dementia, Bacterial Meningitis, Alzheimer's Disease, Cerebral Hemorrhage, Cerebral Injury, Neurofibromatosis, Epilepsy, Cerebral Infarction, Migraine
Gastroenterology	H. Pylori Infection, Chronic Gastritis, Gastric Ulcer, Hemorrhoids, GERD, IBS, Acute Gastroenteritis, Gastric Ptosis, Indigestion, Fatty Liver, Anal Fistula, Perianal Abscess, Anal Fissure, Rectal Prolapse, Rectal Polyp, Constipation, Liver Dysfunction
Pulmonology	Asthma, Pulmonary Tuberculosis, Lung Cancer, Emphysema, Bronchitis, COVID-19 (Novel Coronavirus), Pneumonia, Influenza A, Influenza B, Mycoplasma Infection
Psychiatry	Depression, ADHD, Auditory Hallucination, Anxiety Disorder, Schizophrenia, OCD, Insomnia, Bruxism, PTSD, Dissociative Disorders, Paranoid Personality Disorder, Social Anxiety Disorder, Anorexia, Eating Disorders, Phobia, Bipolar Disorder, Delusional Disorder

Table 5: **Disease Taxonomy.** The dataset spans 98 distinct disease types across six major clinical specialties.

text memory without structured decomposition. **2) Static Structured Schema:** The agent uses structured records but lacks dynamic episodic updates. **w/o Dual-Track Cognitive Memory:** The full framework but with the *NLI-Verifier* loop removed, accepting all generated responses without consistency checks.

Baselines for Downstream Clinical Utility (RQ3) To assess the utility of our synthetic data in training downstream clinical reasoning models, we compare against leading open-source medical LLMs and real-data training setups:

- **State-of-the-art Medical LLMs: 1) Med42-v2-8B** (Christophe et al., 2024): A specialized clinical LLM fine-tuned on extensive medical instruction datasets, serving as a strong domain-specific baseline. **2) UltraMedical-8B** (Zhang et al., 2024): A state-of-the-art medical model enhanced with large-scale biomedical literature and clinical guidelines. **3) Baichuan-M2-32B** (Team et al., 2025): A large-scale bilingual medical foundation model demonstrating superior performance on clinical reasoning tasks.
- **Control Group Settings: 1) Qwen3-8B (Base Model)** (Yang et al., 2025): The unmodified foundation model, serving as the zero-shot baseline to measure the raw capability gain. **2) Qwen3-8B + Real Data:** The base model fine-tuned on real-world clinical records under identical experimental settings. This serves as the *ceiling performance* reference to verify if synthetic data can match the supervision quality of organic data.

C.2 Evaluation Metrics Definition

We provide detailed definitions for the metrics used to evaluate data quality and interaction fidelity.

(RQ1) Data Quality Assessment To validate the synthesized patient records $\mathcal{P}$, we utilize three categories of metrics:

- **Linguistic Quality:** We measure the fluency and lexical richness of the generated text. *Perplexity (PPL)* quantifies the model's uncertainty, with lower scores indicating greater fluency. Instead of using generic open-domain models, we employ a state-of-the-art specialized medical LLM, **Baichuan-M2-32B** (Team et al., 2025) to ensure scores reflect clinical semantic coherence rather than generic syntactic probability. *Distinct-4* calculates the ratio of unique 4-grams, serving as a proxy for lexical diversity and the absence of repetitive degeneration. This ensures that the perplexity score reflects **clinical semantic coherence** and domain-specific plausibility rather than mere syntactic probability. A lower PPL under this model indicates that the generated record aligns well with professional medical corpus distributions.
- **Semantic Diversity:** To ensure **Patient-Zero**

covers the "long tail" of patient distributions without mode collapse, we report *Self-BLEU*, where lower scores indicate higher diversity between generated samples. Additionally, *Entity Diversity* measures the unique count of medical entities recognized by a biomedical NER tagger, reflecting the clinical richness of the dataset.

- **Clinical Validity:** We employ GPT-4.1 (version gpt-4.1-2025-04-14) as a verified evaluator to assess medical alignment. *Consistency* measures the logical coherence between symptoms, medical history, and diagnosis. *Completeness* evaluates whether the record contains all necessary clinical sections (e.g., HPI, Physical Exam) required by standard guidelines.

(RQ2) Interaction Fidelity Assessment To evaluate the agent's performance during dynamic consultations, we define metrics across three dimensions aligned with our Dual-Track architecture:

- **Factual Fidelity (Cognitive):** Assesses the stability of the semantic memory $\mathcal{M}_{sem}$. *Logical Consistency* is the percentage of agent responses that do not logically contradict the ground-truth patient record, verified via an NLI model. *Factual Recall* reports the F1 score of key medical facts retrieved during the conversation compared to the static patient record.
- **Behavioral Fidelity (Persona):** Evaluates adherence to the injected persona Ψ. *Persona Alignment* is reported as a normalized percentage score, measuring the extent to which the agent's responses exhibit the specific behavioral traits as assessed by expert evaluation. *Stylistic Stability* measures the consistency of linguistic style features across multi-turn interactions.
- **Safety & Robustness:** *Hallucination Rate* quantifies the frequency of fabricating non-existent symptoms or medical history. *Inducibility Resistance* measures the agent's safety capability to reject "leading questions" from the doctor (e.g., coercing the patient to admit a symptom they do not possess), which is critical for preventing misdiagnosis in medical simulation.

C.3 Downstream Training

Framework and Hardware We conducted all downstream training experiments on a computational node equipped with 8 NVIDIA A100 80GB GPUs. Our implementation is built upon the VERL framework (Sheng et al., 2025), utilizing efficient distributed training strategies. To optimize inference throughput during the rollout phase, we integrated the vLLM engine. The policy optimization was performed using the Group Relative Policy Optimization (GRPO) (Shao et al., 2024) algorithm, a state-of-the-art method for stabilizing reinforcement learning in language models.

Model Configurations We utilized the Qwen3-8B (Yang et al., 2025) instruction-tuned model as the backbone for both the Actor (Policy) and Critic networks. To manage memory efficiency while maintaining precision, the models were loaded in `bfloat16` format. Gradient checkpointing was enabled to reduce memory footprint. For the optimization process, we employed separate learning rates for the actor and critic to ensure stable convergence: the Actor was trained with a learning rate of 1×10^{-6}, while the Critic utilized a higher learning rate of 1×10^{-5}. We set the KL divergence coefficient (β_{KL}) to 0.001 to prevent the policy from deviating excessively from the reference model.

Training Data and Hyperparameters The training was conducted on the **Patient-Zero** synthetic dataset. We configured a global training batch size of 128, distributed across GPUs with a micro-batch size of 4 per device. The generation configuration allowed for a maximum prompt length of 2048 tokens and a maximum response length of 2048 tokens, ensuring sufficient context window for complex clinical reasoning. During the rollout phase, we sampled $n = 5$ trajectories per prompt to robustly estimate the advantage function. The entire training process was monitored using Weight & Biases (Wandb)[3], with a reward clipping mechanism applied to maintain training stability.

Benchmark To rigorously assess whether the specialized training on synthetic patient data induces *catastrophic forgetting* of general medical knowledge, we extended our evaluation to two gold-standard multiple-choice benchmarks. The evaluation protocol was standardized on a randomly sampled subset of 200 questions from the **US English** test sets for MedQA dataset (Jin et al., 2020) and MMLU (Hendrycks et al., 2021) of medical topics.

D Statistical Distribution Analysis

To rigorously validate the claim of *Real-World Distribution Alignment*, we define the statistical framework used for the **Chi-square Goodness-of-Fit**

[3] https://wandb.ai/site

Table 6: Hyperparameters for Downstream Training.

Hyperparameter	Value
General Settings	
Base Model	Qwen3-8B
Algorithm	GRPO
Precision	bfloat16
Number of GPUs	8
Batch & Rollout	
Global Batch Size	128
PPO Mini-Batch Size	64
Micro-Batch Size (Per GPU)	4
Rollout Samples (N)	5
Optimization	
Actor Learning Rate	1×10^{-6}
Critic Learning Rate	1×10^{-5}
KL Coefficient (β_{KL})	0.001
Max Prompt Length	2048
Max Response Length	2048

test (χ^2). We compared the generated **Patient-Zero** cohort ($N = 200$) against ground-truth epidemiological priors derived from our Knowledge Base (Monte Carlo reference, $N = 2000$).

Definition D.1 (Null Hypothesis for Distribution Alignment)**.** Let $\hat{\mathcal{P}}$ be the empirical distribution of the synthetic patient attributes and $\mathcal{P}^*$ be the real-world prior distribution. The null hypothesis H_0 posits that there is no statistically significant difference between the two distributions:

$$H_0 : \hat{\mathcal{P}} = \mathcal{P}^*$$

We adopt a significance level of $\alpha = 0.05$. A p-value > 0.05 indicates a failure to reject H_0, confirming alignment.

Hypothesis Testing Results Table 7 presents the results of the Chi-square Goodness-of-Fit tests across nine epidemiological features. We observe that **Patient-Zero** successfully passed the alignment test for all nine features (100% success rate). Notably, complex lifestyle attributes, which are typically prone to generation bias in LLMs, showed exceptionally high conformity: *Alcohol Consumption* ($p = 0.97, \chi^2 = 0.07$) and *Dietary Habits* ($p = 0.71, \chi^2 = 0.69$). The *Fisher's combined probability test* yielded a global p-value of **0.86**, providing strong evidence that the synthetic cohort is statistically indistinguishable from the real-world population.

Effect Size Analysis To ensure that the lack of statistical significance is due to genuine distributional similarity rather than insufficient sample power, we computed effect size metrics including *Cramér's V* and *Total Variation Distance (TVD)*.

- **Cramér's V:** The average Cramér's V across all features is **0.054**, with a maximum of 0.077 (Physical Activity). According to Cohen's guidelines (Cohen, 1988), a value $V < 0.10$ indicates a ***negligible effect size***, confirming that the deviation between synthetic and real distributions is minimal.

- **Total Variation Distance (TVD):** The average TVD is **0.035**, implying that the synthetic probability mass function deviates from the ground truth by an average of only 3.5% across all categories.

Granular Category Alignment We further examined the alignment at the granular category level. For instance, in the *Ethnicity* feature, the synthetic distribution for minority groups closely mirrors the expected priors (e.g., Hispanic: Observed 17.0% vs. Expected 14.3%, $\Delta = +2.7\%$). Similarly, in *Socioeconomic Status*, the Low Income group representation (39.0%) is nearly identical to the ground truth (39.9%). This granular fidelity ensures that **Patient-Zero** preserves the diversity of underrepresented subpopulations, mitigating the risk of mode collapse often seen in synthetic medical data.

E Human Expert Evaluation Details

E.1 Evaluation Protocol

We recruited two senior licensed physicians from top-tier hospitals to conduct a blinded evaluation. The evaluation consisted of two distinct tasks:

- **Task 1: Human-AI Data Discrimination Test.** Experts were presented with a random mix of 28 anonymized patient records and were asked to classify the origin of each record as either "Human-authored" or "AI-generated".

- **Task 2: Multi-dimensional Quality Assessment.** Experts rated 30 distinct records on a 5-point Likert scale across four clinical dimensions: **1) Linguistic Quality:** Fluency, professional tone, and absence of redundancy. **2) Diversity (Information Richness):** Depth of medical details (e.g., specific symptom attributes, history). **3) Consistency:** Logical coherence between symptoms, diagnosis, and treatment.**4)**

Table 7: **Chi-Square Goodness-of-Fit Test Results.** The high p-values across all dimensions (especially Lifestyle attributes like Alcohol Consumption) strongly support the acceptance of H_0.

Dimension	Feature	χ^2 Statistic	p-Value	Result
Demographics	Age Group	0.82	0.66	Aligned
	Biological Sex	1.00	0.32	Aligned
	Ethnicity	3.40	0.49	Aligned
Socioeconomic	Income Level	0.62	0.73	Aligned
	Geographic Location	1.37	0.51	Aligned
Lifestyle	Smoking Status	2.21	0.33	Aligned
	Alcohol Consumption	**0.07**	**0.97**	Aligned
	Physical Activity	2.35	0.31	Aligned
	Dietary Habits	0.69	0.71	Aligned
Fisher Combined Statistic		-	***0.86***	***Global Match***

Table 8: **Descriptive Statistics of Expert Likert Ratings.** Patient-Zero outperforms Real Data across multiple dimensions.

Source	Metric	N	Mean	SD	Median	95% CI
Patient-Zero	Linguistic Quality	10	2.40	0.52	2.0	[2.03, 2.77]
	Information Richness	10	3.40	0.70	3.5	[2.90, 3.90]
	Consistency	10	3.30	0.82	3.5	[2.71, 3.89]
	Completeness	10	3.00	1.16	3.5	[2.17, 3.83]
Real Data	Linguistic Quality	10	1.80	1.14	1.0	[0.99, 2.61]
	Information Richness	10	1.90	1.10	1.5	[1.11, 2.69]
	Consistency	10	3.30	0.95	4.0	[2.62, 3.98]
	Completeness	10	2.20	0.92	2.0	[1.54, 2.86]

Completeness: Presence of all necessary clinical sections (e.g., Chief Complaint, HPI, Physical Exam).

E.2 Statistical Analysis of Expert Ratings

We performed pairwise Mann-Whitney U tests to quantify the differences between data sources. Table 8 details the descriptive statistics and significance tests.

Performance against Real Data Real Data received the lowest scores in *Linguistic Quality* (Mean=1.80, SD=1.14) and *Completeness* (Mean=2.20, SD=0.92), reflecting the inherent noise in real-world clinical notes (e.g., abbreviations, typos, missing sections). In contrast, **Patient-Zero** achieved significantly higher scores than Real Data in *Information Richness* ($U = 85.5, p = 0.006$, Cohen's $d = 1.63$) and comparable scores in other metrics. This statistical evidence confirms that our synthetic data not only matches but in specific aspects surpasses the quality of raw human-authored records.

Rater Reliability Analysis To ensure the objectivity of the human evaluation, we examined the scoring distribution differences between the annotators. A *Kruskal-Wallis test* revealed no statistically significant difference in their rating patterns ($H = 0.054, p = 0.8167$). The high p-value indicates strong inter-rater consistency, confirming that the observed performance gaps are attributable to model quality rather than annotator bias.

E.3 Qualitative Case Studies

To understand the specific strengths and weaknesses of each source, we analyzed the qualitative feedback provided by the physicians.

Critique of Real Data Physicians frequently noted that real records were chaotic or incomplete. For example, one evaluator commented on a real record: *"Treatment plan is very chaotic... stopped drug for 3 days but didn't say which drug."* Another noted: *"Incomplete examination results."* These comments validate our premise that raw real-world data often lacks the structural integrity required for high-quality model training.

Critique of Baselines (MERA-Qwen) The baseline model suffered from logical repetition and

translation artifacts. Evaluators noted: *"Past history has repetition"* and *"Diagnosis has Broadwater prefix, likely a translation error."*

Patient-Zero Performance Our framework produced the most idealized records. Physicians explicitly evaluated the generated cases as *"overall complete and accurate"*, noting that critiques were mostly limited to minor omissions rather than logical contradictions. For instance, one comment pointed out a specific missing detail: *"No description of fever in the disease course."* This contrast suggests that **Patient-Zero** successfully captures the high-level structure and logic of clinical narratives, ensuring holistic validity with only fine-grained details occasionally requiring refinement.

F Case Studies and Prompts

[Synthetic Case 1] Judged as Human-authored by Experts: Neurofibromatosis Type 1

Patient Profile

- **Name:** Claire Thompson
- **Demographics:** 54-year-old Female, Caucasian
- **Socioeconomic:** Suburban residence, Low Income
- **Record ID:** 398899a3-8d31-4805-8279-4361209810d9

Presenting Complaint: Cutaneous Lesions & Spinal Pain

- **Symptoms:** Café-au-lait spots (since childhood), Axillary freckling, Cutaneous neurofibromas (slow increase since 20s), Dome-shaped soft nodules, Intermittent spinal pain.
- **History:** Lesions largely stable; mild surge during pregnancies. Spinal pain worse with sitting, relieved by stretching.
- **Family History:** Father with NF1; Mother with Hypertension.

Medical History

- **Diagnosis:** NF1 diagnosed at age 28 ($\geq$6 café-au-lait macules + freckling).
- **Comorbidities:** Mild thoracic scoliosis
  [Image of thoracic scoliosis X-ray] (adolescence), Hypertension risk (monitored).
- **Interventions:** Conservative dermatology follow-up; no prior surgeries.

Examination Results

- **Genetics:** Targeted next-generation sequencing of NF1/NF2 (mean on-target coverage 312x; 99.2% of coding bases >50x) identified a heterozygous truncating variant in NF1: c.2041C>T (p.Arg681*), variant allele fraction 47% in peripheral blood leukocytes. No pathogenic or likely pathogenic variants were detected in NF2. Copy number analysis showed no exon-level deletions/duplications in NF1 or NF2. Sanger confirmation of the NF1 variant was positive. Microarray (SNP/CGH) was unremarkable.
- **Ultrasound:** High-frequency (15 MHz) soft tissue ultrasound of trunk and upper arms: 14 discrete subcutaneous nodules, ovoid, well-circumscribed, hypoechoic relative to surrounding fat (mean echogenicity 19–23 grayscale units vs fat 34–38), located within dermis/subcutis at depths 2–6 mm. Largest lesion: right posterior shoulder 1.2 x 0.9 x 0.7 cm; majority 0.4–0.8 cm. Lesions are compressible and mobile over deep fascia; no deep fascial involvement. Color Doppler shows minimal internal vascularity in 3/14 lesions (scattered low-flow signals; peak systolic velocity 4–6 cm/s; resistive index 0.56–0.62). Axillary region: skin freckling noted externally; no discrete mass on sonography. Focused abdominal and pelvic survey: liver (size 15.8 cm, homogeneous), spleen (10.9 cm), kidneys (R 10.8 cm, L 10.6 cm), uterus and adnexa within normal sonographic limits; no intra-abdominal solid masses.
- **MRI Examination:** Thoracic spine MRI (1.5T): At right T9 dorsal root region, an ovoid lesion measuring 7 x 6 x 5 mm along the nerve sheath. T1-weighted images: isointense to skeletal muscle (signal intensity ratio lesion/muscle 1.05). T2-weighted images: markedly hyperintense with a central slightly low-signal area (target sign) occupying 42% of the cross-sectional area. STIR: strong hyperintensity. Post-gadolinium T1: mild heterogeneous enhancement (signal increase 38%). No spinal cord edema, no mass effect; canal remains patent. Additional findings: mild degenerative changes with L4–L5 posterior disc protrusion 2 mm, no nerve root contact; thoracic coronal localizer shows mild scoliosis measuring 11°.
- **Spinal Angiography:** Selective catheter spinal angiography: right T9 intercostal/radiculomedullary artery injection demonstrates normal arterial phase and opacification of the anterior spinal artery without arteriovenous shunting. A subtle contrast column indentation is observed at the right nerve root exit zone, consistent with a small extrinsic filling defect. No abnormal tumor blush; contrast remains intrathecal without extravasation. No accidental contrast injection into a mass. Other levels show normal vascular anatomy.
- **Ophthalmic Examination:** Best-corrected visual acuity: OD 20/25, OS 20/30 (with +1.50 D add for near). Intraocular pressure: OD 14 mmHg, OS 13 mmHg (Goldmann applanation). Slit-lamp: multiple raised, smooth, brown-yellow iris nodules consistent with hamartomas-OD: 6 nodules ranging 0.8-1.5 mm; OS: 5 nodules ranging 0.7-1.3 mm; located predominantly in inferior and temporal quadrants. Cornea and anterior chamber clear. Optic discs sharp with cup-to-disc ratio 0.3 OU; no edema, pallor, or atrophy. Optical coherence tomography RNFL: average thickness OD 99 μm, OS 97 μm; macular cube thickness OD 279 μm, OS 276 μm. Humphrey 24-2 visual fields: MD OD -0.8 dB, OS -1.1 dB; pattern deviation without focal defects.
- **X-ray Examination:** Standing PA and lateral thoracolumbar radiographs: mild right-convex thoracic scoliosis from T6–T11 with Cobb angle 12°. Vertebral body heights maintained; pedicles symmetric; no ribbon-like lucencies or longitudinal stripe-like lesions within bones. Mild multilevel spondylotic changes (small anterior osteophytes at T8–T10 and L4–L5). No vertebral scalloping, no dysplasia of long bones, ribs, or pelvis. Overall skeletal involvement minimal on plain films.

Table 9: **Synthetic Patient Record (Neurofibromatosis Type 1).** This record was judged as ***human-authored"*** by the licensed physician. It presents a classic, longitudinal clinical picture of Neurofibromatosis Type 1 in a middle-aged female, capturing the typical evolution of cutaneous markers (pregnancy-associated progression), specific ophthalmological criteria (Lisch nodules), and characteristic radiological findings (Target sign on MRI).

[Synthetic Case 2] Judged as Human-authored by Experts: Pneumonia

Patient Profile

- **Name:** Rohan Kumar
- **Demographics:** 21-year-old Male, Asian
- **Socioeconomic:** Rural residence, Middle Income
- **Record ID:** ef536bce-a379-4a50-9140-d889740195bd

Presenting Complaint: Respiratory Infection (Mild Severity)

- **Symptoms:** Productive cough (yellow purulent sputum), low-grade fever, mild pleuritic chest pain, fatigue, loss of appetite.
- **History:** Cold-like onset 5 days ago; progressed to productive cough. Stable/improving with OTC care.
- **Risk Factors:** Current smoker (7 cigs/day), Indoor wood-smoke exposure.

Medical History

- **Comorbidities:** Seasonal allergic rhinitis. History of acute bronchitis (age 19).
- **Lifestyle:** Vegetarian diet. Heavy alcohol use (weekend binge).

Examination Results

- **Complete Blood Count:** Collected at 08:20 on day 5 of illness. WBC $11.6x10^9/L$ (ref 4.0–10.0), with neutrophils 78% (absolute neutrophil count $9.0x10^9/L$), lymphocytes 15% $(1.7x10^9/L)$, monocytes 5% $(0.58x10^9/L)$, eosinophils 1% $(0.12x10^9/L)$, basophils 1% $(0.12x10^9/L)$. Bands 3% noted. Hemoglobin 14.3 g/dL, hematocrit 42.1%, MCV 87 fL, MCH 29 pg, RDW 12.9%. Platelets $325x10^9/L$. Findings indicate mild leukocytosis with neutrophil predominance.
- **C-Reactive Protein and Procalcitonin:** Serum CRP 38 mg/L (ref <5 mg/L). Procalcitonin 0.12 ng/mL (ref <0.05 ng/mL; values <0.25 ng/mL generally associated with milder bacterial activity). Repeat CRP at 48 hours planned; current trend reported by patient (symptomatic) is slightly improving with antipyretics.
- **Oxygenation Assessment and Arterial Blood Gas Analysis:** Vital signs at rest (room air, FiO2 0.21): Temp 37.9°C, HR 88 bpm, RR 18/min, BP 118/72 mmHg. Pulse oximetry SpO2 97–98% at rest; with brisk 3-minute walk SpO2 95–96% and RR 20/min, recovery to 97% within 1 minute. Arterial blood gas (right radial artery, room air): pH 7.44, PaCO2 36 mmHg, PaO2 92 mmHg, HCO3- 24 mmol/L, Base excess +0.5 mmol/L, SaO2 97%. Calculated alveolar-arterial (A–a) O2 gradient 13 mmHg (slightly above age-adjusted expected). Lactate 1.2 mmol/L (ref 0.5–2.2). Carboxyhemoglobin 2.1% (consistent with current smoker), Methemoglobin 0.5%. Overall oxygenation adequate without resting hypoxemia.
- **Sputum Smear, Sputum Culture, and Drug Sensitivity Test:** Early-morning expectorated sputum, 3 mL, yellow mucopurulent. Gram stain: >25 polymorphonuclear leukocytes/LPF, <10 squamous epithelial cells/LPF (good-quality specimen). Predominant gram-positive lancet-shaped diplococci observed; occasional gram-negative rods; no acid-fast bacilli; fungal elements not seen. Culture (48 hours): Moderate growth (2+) of alpha-hemolytic colonies on 5% sheep blood agar; optochin zone 20 mm and bile solubility positive, consistent with Streptococcus pneumoniae. Normal respiratory flora also present; no growth of Staphylococcus aureus or Pseudomonas aeruginosa. Antimicrobial susceptibility (MIC, CLSI categories): Penicillin (parenteral) 0.06 mg/L – S; Amoxicillin/clavulanate 0.5 mg/L – S; Ceftriaxone 0.25 mg/L – S; Levofloxacin 1 mg/L – S; Erythromycin 8 mg/L – R; Azithromycin 8 mg/L – R; Doxycycline 2 mg/L – I; Trimethoprim-sulfamethoxazole >4 mg/L – R; Vancomycin 0.5 mg/L – S; Linezolid 1 mg/L – S.
- **Imaging Examination:** Chest X-ray (PA and lateral) performed on day 5 of illness. Cardiomediastinal silhouette normal; cardiothoracic ratio 0.49. Lungs: patchy air-space opacity in the right lower lobe, posterior basal segment, measuring approximately 3.2 x 2.1 cm, with faint air bronchograms and mild peribronchial thickening in the right infrahilar region. No pleural effusion; costophrenic angles sharp. No cavitation or pneumothorax. Lateral view localizes opacity to the posterior segment of the right lower lobe. Findings consistent with a focal infectious/inflammatory process of limited extent.

Table 10: **Synthetic Patient Record (Pulmonology).** This record was judged as judged as ***human-authored*** by licensed physician. It accurately simulates a typical mild Community-Acquired Pneumonia (CAP) in a young adult, featuring the classic progression from viral prodrome to bacterial superinfection, coherent microbiological findings (pneumococcal diplococci), and consistent radiographic evidence (lobar consolidation).

Rejected Sample Analysis: 76-year-old Male of Colorectal Cancer

Patient Profile

- **Name:** Arthur Whitaker
- **Demographics:** 76-year-old Male, Caucasian
- **Socioeconomic:** Rural residence, Low Income
- **Target Severity Label: Mild**

Synthesized Clinical Narrative (Draft)

- **Symptoms:** Change in bowel habits, bloody stool, abdominal pain, diarrhea. Notably reports **progressive fatigue**, **unintentional weight loss (3kg)**, and **loss of appetite**.
- **History:** Symptoms began 3–4 months ago. Initial rectal bleeding followed by constitutional symptoms. Lab work revealed microcytic iron-deficiency anemia.
- **Diagnosis:** Colonoscopy confirmed localized left-sided adenocarcinoma.

Rejection Logic

- **Conflict Type:** Severity Mismatch.
- **Reasoning:** The generated narrative includes significant constitutional symptoms (*weight loss, anorexia, fatigue*) and systemic complications (*iron-deficiency anemia*). These clinical features indicate a disease burden that exceeds the clinical definition of "Mild" for early-stage colorectal cancer, which is typically asymptomatic or presents with minor local symptoms only.
- **Outcome:** ***Rejected*** by NLI-Verifier. Triggered regeneration with strict constraints on systemic symptoms.

Table 11: **Quality Control Visualization.** An example of a synthetic record rejected by our framework. The *NLI-Verifier* detected a semantic contradiction where the generated symptom burden (systemic constitutional symptoms) conflicted with the input condition of "Mild" severity, ensuring strict clinical consistency in the final dataset.

Rejected Sample Analysis: 44-year-old Female of Cirrhosis

Patient Profile

- **Name:** Layla Al-Khatib
- **Demographics:** 44-year-old Female, Middle Eastern
- **Socioeconomic:** Suburban residence, Low Income
- **Target Condition: Compensated Cirrhosis (Mild)**

Synthesized Clinical Narrative (Draft)

- **Medical History:** History of alcohol use disorder. Compensated cirrhosis diagnosed 1 year ago based on ultrasound showing coarse liver and mild splenomegaly **without ascites**.
- **Generated Symptoms:** Fatigue, loss of appetite, weight loss, and **abdominal distension**.
- **Narrative Detail:** "She notes mild **abdominal fullness/distension** for the past 6 months... No episodes of jaundice or ascites to date."

Rejection Logic

- **Conflict Type:** Logical Contradiction (Pathophysiology).
- **Reasoning:** The generated narrative lists "abdominal distension" as a symptom while explicitly stating the patient has "no ascites" and is in the "compensated" stage. In cirrhosis, abdominal distension is primarily caused by fluid accumulation (ascites). Asserting distension in the absence of ascites (or gas/obesity, which were not contextualized) creates a medically incoherent presentation for a liver disease case.
- **Outcome:** ***Rejected*** by NLI-Verifier. The record was flagged for factual inconsistency regarding signs of decompensation.

Table 12: **Logic Consistency Check.** In this rejected sample, our framework detected a conflict between the generated symptom (*abdominal distension*) and the clinical constraints (*no ascites, compensated stage*), preventing the generation of physiologically impossible patient records.

Prompt Template: Medically-Aligned Disease Outline Generation

System Instruction

- **Role:** Senior Medical Expert.
- **Task:** Analyze raw disease info and generate a refined JSON outline for synthetic record generation.
- **Output Format:**
  STRICT JSON format containing: `disease_summary`, `key_characteristics`, `typical_presentation` (mild/moderate/severe), `important_notes`, `contraindications`, `differential_considerations`, `special_populations`, and `red_flags`.

User Prompt Template

- **Context Injection:** Analyze the following information:
  - `{raw_outline}`: [Raw medical knowledge text]
  - `{profile_context}`: [e.g., Demographics, Risk Factors]
  - `{severity_context}`: [e.g., Target Severity Level]

Generation Instructions & Constraints

1. Summarize key clinical characteristics concisely.
2. Describe typical presentations across severity levels (Mild, Moderate, Severe).
3. List **Important Notes** for generation: Age/Gender manifestations, symptom timing, comorbidities.
4. Identify **Contraindications**: Incompatible demographics, conflicting symptoms, unrealistic findings.
5. Note considerations for **Special Populations** (Pediatric, Elderly, Pregnant).
6. List **Red Flags** indicating severe complications.

Table 13: **Disease Outline Generation Prompt.** The prompt guides the LLM to structure unstructured medical knowledge into a hierarchical JSON format, enforcing clinical constraints and severity-specific presentations before the actual patient record generation begins.

Prompt Template: Patient Record & Symptom Generation

System Instruction

- **Role:** Knowledgeable Medical Expert.
- **Task:** Generate a comprehensive patient record with realistic symptoms based on the provided profile.

Input Data Injection

- **Target Condition:** `{disease_name}`, `{severity_level}`
- **Demographic Profile:**
  - Biological: `{Age}`, `{Sex}`, `{Physiological_Status}`
  - Sociocultural: `{Ethnicity}`, `{Geography}`, `{Socioeconomic}`
  - Behavioral: `{Smoking}`, `{Alcohol}`, `{Diet}`, `{Lifestyle}`
- **Clinical Knowledge Base:**
  - `{symptoms_list}`: List of possible symptoms for this disease.
  - `{disease_outline}`: Structured characteristics and progression logic.

Generation Directives

1. **Identity Synthesis:** Generate a realistic name reflecting the patient's ethnicity and country of origin.
2. **History Construction:** Create a detailed medical history consistent with the disease and age profile.
3. **Background Factors:** Synthesize lifestyle factors, vaccination history, and family history.
4. **Symptom Selection:** Select specific symptoms from the provided list appropriate for the target `severity`.
5. **Chronology:** Specify symptom duration and progression logic.

Output Requirement: Output ONLY a valid JSON object.

Table 14: **Patient Record Generation Prompt.** This core module synthesizes the abstract demographic parameters (e.g., "Middle Income", "Smoker") and medical knowledge into a concrete patient identity (e.g., "Name", "Specific History"), ensuring the resulting profile is culturally consistent and clinically plausible.

Prompt Template: Examination Result Generation

System Instruction
- **Role:** Knowledgeable Medical Expert.
- **Task:** Generate specific, quantitative examination findings based on patient symptoms and disease severity.

Input Data Injection
- **Case Context:**
  - Patient: `{patient_info}`, `{symptoms_data}`
  - Disease: `{disease_name}`, `{severity_level}`
- **Medical Grounding:**
  - `{exam_list}`: List of required examinations (e.g., CBC, CT Scan).
  - `{exam_reference}`: **Reference standards** and normal ranges for each exam.
  - `{disease_outline}`: Pathophysiological characteristics guide.

Generation Directives
1. **Quantitative Specificity:** Include realistic values, percentages, and measurements (e.g., "Hemoglobin 8.5 g/dL" rather than "Low Hemoglobin").
2. **Severity Alignment:** Findings must strictly reflect the `{severity}` (e.g., Mild fibrosis vs. Decompensated ascites).
3. **Diagnosis Blinding: Do NOT** directly mention the disease name in the results. Describe the *signs*, not the *label*.
4. **Clinical Realism:** Mimic the observation style of real clinical reports.

Output Requirement: Output ONLY a valid JSON object.

Table 15: **Examination Generation Prompt.** Unlike open-ended generation, this module is grounded by injected `Reference Standards`, forcing the LLM to generate precise numerical data and observations that are statistically plausible within the specific clinical context.

Prompt Template: Clinical Symptom and Consistency Validation

System Instruction
- **Role:** Medical Expert Reviewer.
- **Task:** Review a generated patient record for clinical accuracy and guideline compliance.

Input Data Injection
- **Disease Metadata:** `{disease_name}`, `{severity_level}`
- **Patient Profile:**
  - Age: `{age}` (`{age_group}`)
  - Sex: `{sex}`
  - Physiological Status: `{physiological_status}`
  - Smoking Status: `{smoking_status}`
- **Generated Record:** `{json.dumps(response)}`

Validation Criteria (Clinical Guidelines)
1. **Symptom Consistency:** Symptoms must be clinically consistent with the disease and severity.
2. **Severity Matching:** Symptom count/intensity must match the stated level (Mild vs. Severe).
3. **Demographic Logic:** Age/Sex must be logically consistent with medical history.
4. **Contradiction Check:** No impossible combinations (e.g., pregnant male, pediatric conditions in elderly).
5. **Realistic Progression:** Duration description should be realistic for the disease.
6. **Plausible History:** Medical history must align with age and condition.

Output Requirement
- Evaluate strict compliance with the above guidelines.
- **Format:** Output ONLY a JSON object.

Table 16: **Symptom Validation Prompt.** This prompt acts as an automated critic, employing a frozen LLM to verify that the generated patient narrative strictly adheres to the input constraints and clinical pathophysiology before the record is accepted into the dataset.

Prompt Template: Examination Results Verification

System Instruction
- **Role:** Medical Expert Reviewer.
- **Task:** Review generated examination results for clinical accuracy, realism, and guideline compliance.

Input Data Injection
- **Clinical Context:** `{disease_name}`, `{severity_level}`
- **Patient Data:** `{patient_info}`, `{symptoms_data}`
- **Target:**
  - Generated Results: `{exam_results}`
  - Expected Standard: `{expected_exams}`

Validation Criteria (Clinical Guidelines)
1. **Clinical Consistency:** Results must align with the disease pathology and reported symptoms.
2. **Realistic Values:** Measurements and observations must be within plausible biological ranges.
3. **Severity Reflection:** Results should appropriately reflect the stated severity (e.g., mild vs. severe abnormalities).
4. **Findings Only:** Results must NOT directly state the diagnosis (describe the *signs*, not the *conclusion*).
5. **Completeness:** All critical examinations for the specific condition must have meaningful entries.
6. **Internal Logic:** No contradictory findings between different test modalities (e.g., Lab vs. Imaging).
7. **Unit Precision:** Numeric values must use correct medical units.

Output Requirement
- Evaluate strict compliance with the above guidelines.
- **Format:** Output ONLY a JSON object.

Table 17: **Examination Validation Prompt.** This module ensures that generated medical test results (e.g., Labs, Imaging) are not only clinically accurate but also statistically realistic, preventing "diagnosis leakage" where the model prematurely states the conclusion instead of raw findings.