COHORTWORLD: ALIGNING REAL-WORLD DATA-DRIVEN DIAGNOSTICS AND LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical language models (LMs) have shown great potential in complex real-world healthcare tasks since they can act as agents in interactive virtual clinical environments with verifiable rewards as trustworthy feedback. However, the accuracy of data-driven diagnostics in clinic practice is challenged by label quality due to restrictive data use agreements, synthetic yet access-gated health environment, and overly narrow health context, which leads to the absence of real-world evidence (RWE) verifying LMs to avoid hallucination in medical LMs. To bridge this gap, we present *CohortWorld*, an *in-silico* environment generated from large-scale UK Biobank (UKB) cohort in which each state is a real-world cohort anchored in disjoint histories of disease, procedure, and medication. *CohortWorld* is an reinforcement learning (RL) suite based on disjoint anonymous cohorts (n=29,884) with real-world data gathered from >500k UKB subjects for evaluating LMs on disease diagnosis and simulating RWE reasoning as a human doctor. Our empirical testing demonstrates that existing closed- and open-source LMs severely underperform on the RWE-calibrated benchmark, hovering at or below an AUROC of 0.5 with suboptimal slope metrics for disease risk ranking. In contrast, Qwen3.5-9B outperforms Sonnet 5 using our *CohortWorld* framework, demonstrating the benefits of RWE matching and counterfactual risk modification. Together, *CohortWorld* advances the trustworthiness of artificial diagnosis and enables LMs to reason like a human doctor based on verifiable RWE in a releasable and actionable *in-silico* environment.

1 INTRODUCTION

Beyond simple question answering, LM agents are growing in adoption for complex real-world workflows, a transition enabled by their notable achievements across mathematics Trinh et al. (2024), gaming Wang et al. (2023), code generation Jimenez et al. (2024), and OS-level computer operation Xie et al. (2024). In this context, patients start to seek an AI-empowered diagnosis from LMs prior to the real diagnosis from human doctors, yet the LM still hallucinates in diagnostics, leading to disagreement with experienced doctors Singhal et al. (2025) (Fig. 1a). Although LMs have become capable at mathematics and programming inside interactive environments whose feedback is verifiable Wen et al. (2026), medical LMs in diagnostics have not been fully verified or benchmarked.

The medical history of diagnoses World Health Organization (2004) (2004), treatment regimens Joint Formulary Committee (2026); Connecting for Health (Organization) (2009), and survival outcomes AbouZahr et al. (2010) has been documented in electronic health records (EHR). However, because these records involve identifiable individuals, they cannot be publicly distributed. There are informed consent and data use agreements (DUAs) to strictly control access. Existing medical LM benchmarks avoid publishing raw records in two ways. First, synthetic environments simulate medical interactions from clinical dynamics so that decision-making can be benchmarked without real-patient risk Luo et al. (2024); Kuo et al. (2022). Second, public exam resources score closed-ended knowledge retrieval and reasoning on professional board questions Yang et al. (2025); Cai et al. (2024). Synthetic environments ground their rewards in a generative model instead of real-world data, and exam benchmarks ground their answers in a curated hypothetical patient instead of a real human, so neither can verify the trustworthiness of diagnostics nor expose the hallucinated risk that requires real-world evidence (RWE) to reject.

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

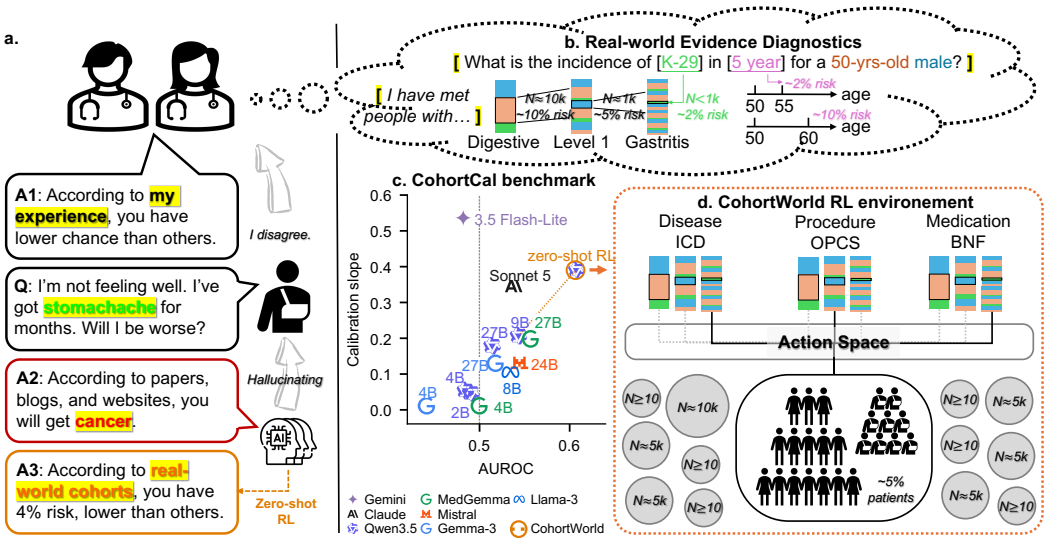

Figure 1: **(a)** Patients seek an artificial diagnosis from LMs prior to the real diagnosis from human doctors. **(b)** The reasoning path of diagnosis by doctors is bracketed by '[]' and based on the experience of building real-world evidence (RWE) and medical knowledge. **(c)** A new benchmark of aligning RWE and LMs for disease risk. **(d)** CohortWorld aligns RWE and LMs by enabling LMs to reason like a human doctor.

RWE is the clinical evidence derived from analyzing EHRs of patients, and it is accepted by the FDA as admissible evidence of effectiveness and safety for diagnostics Schurman (2019), since it reproduces the conclusion of the corresponding randomized trial in a large fraction of replicated cases Franklin et al. (2021). Therefore, RWE diagnostics is a trustworthy practice for anticipating an individual disease risk. As shown in Fig. 1b, a doctor matches the medical history of a patient with the cohorts they have seen in practice, reasons about the differences that remain after matching, and applies the effect of those differences on incidence based on medical knowledge. The example reasoning path in Fig. 1b is a sequence of progressively narrower coding from a general digestion chapter, a level-1 code covering multiple digestive organs, to the target gastritis code in green, with the consideration of various lengths of future horizon (in pink). Following the same reasoning path as a human doctor, we **hypothesize** that if an LM whose diagnostic Chain-of-Thought (CoT) is forced onto RWE, matching a real cohort before reasoning on the relationship between a target disease and medical history, the AI diagnosis report will become more trustworthy.

In this work, we seek to align the aforementioned real-world data-driven diagnostics, a.k.a. RWE diagnostics, and LMs by removing the accessibility barrier and providing an actionable world of cohorts for an LM to reason on RWE. Given the high variability of disease, intervention, and medication history, a population-scale biobank therefore can decompose into a large number of disjoint cohorts, each carrying averaged outcomes. Removing identifiable individual information leaves these cohorts carrying real health records with privacy that is preserved from the public, whereas existing benchmarks Wornow et al. (2023); Jiang et al. (2025) built directly on records remain behind credentialed access due to identifiable individuals. To this end, we introduce CohortWorld, defining its action space through clinician-like patient-cohort matching that enables targeted interventions in patient medical histories. By iterating all possible parameters in the proposed action space over the UK Biobank dataset Sudlow et al. (2015), CohortWorld collects a comprehensive set of n=29,884 downloadable cohorts from >500k subjects, where each cohort is labeled with the incidence of all ICD-10 coded diseases.

Among CohortWorld, the RWE diagnostics benchmark is formed and released as CohortCal for quantifying the alignment between RWE diagnostics and LMs. As shown in Fig. 1c, current closed-source and open-source LMs struggle around or below 0.5 AUROC with a poor calibration slope on disease risk ranking in basic items of CohortCal, where AUROC measures whether a model orders two subjects correctly by their risk of the same disease, and the calibration slope is the product of the spread of risk and the correlation between rankings. Every LM we test lands from the lower-left of the plane with smaller LMs to the upper-right with bigger LMs, following the scaling law Kaplan

et al. (2020). The two closed-source models separate on different axes rather than dominating, where Claude Sonnet 5 is the strongest calibrated model, while Gemini 3.5 Flash-Lite attains the highest slope in the plot at near-chance AUROC, which is the signature of spread without correct ordering. A Qwen3.5-9B given only the CohortWorld harness circled in orange moves up and to the most right, being the single point in the plot that improves discrimination and calibration together. The key advance of CohortWorld lies in formulating RWE diagnostic reasoning as a reinforcement learning (RL) environment (Fig. 1d). By restricting actions to moving between cohorts along the three coding axes for diagnoses (by International Classification of Disease, ICD) World Health Organization (2004) (2004), procedures (by Office of Population Censuses and Surveys, OPCS) Connecting for Health (Organization) (2009), and medications (by British National Formulary, BNF) Joint Formulary Committee (2026), reading real disease risk, and committing estimated risk difference, this environment forces the CoT reasoning path to follow a clinical workflow, which is termed CohortThink in this work, rather than completely relying on medical suggestions by LMs.

Our main contributions are summarized below

1. **RWE diagnostics in an RL environment.** We propose CohortWorld, an interactive environment for language agents whose states are anonymous disjoint cohorts, and actions are diagnostic reasoning steps to look for RWE from cohorts. Together with a new cross-sample GRPO (Group Relative Policy Optimization Shao et al. (2024)), CohortWorld rewards an LM to reason about RWE before diagnosis.
2. **Comprehensive cohort collection.** We propose a greedy granularity ladder that generalizes cohorts across ICD-10, OPCS-4, and BNF coding levels of diagnoses, procedures, and medications, respectively, crossed with coarsened history time ranges, maximizing state-space to 29,884 cohorts and CohortCal benchmark capacity to 70,378 items.
3. **RWE-LM calibration benchmark.** We release an evaluation along with train and test data for two closed- and eleven open-weight LMs on 1,115 distinct disease-risk predictions relative to the population average over CohortCal.

2 Related Works

Medical RL environments. Prior environments differ in how they ground the reward. **Reward from a generative model.** In the DTR-Bench Gymnasium simulator Luo et al. (2024) and the procedurally generated clinical POMDP Hargrave et al. (2024), state transitions and rewards are derived from differential equations and manually specified structural causal models; consequently, policies validated within these environments are tested against assumptions rather than actual empirical evidence. In contrast, Health Gym Kuo et al. (2022) provides data tables fitted using GANs. While these tables contain the necessary environmental variables, they lack defined transition functions or rewards. **Reward from an answer key.** AgentClinic Schmidgall et al. (2024) scores a diagnosis against an examination key voiced by a simulated patient, and MedAgentBench Jiang et al. (2025) scores retrieved answers against physician-written reference solutions. Neither verifies the consequences nor the interventions. **Rewards from observed outcomes.** While offline reinforcement learning for intensive care records Komorowski et al. (2018); Gottesman et al. (2018) and few-shot prediction based on longitudinal records Wornow et al. (2023) preserve real-world outcomes, their environment states correspond to individual patients. Consequently, data access is strictly constrained by credentialing and signed DUAs. Furthermore, the former focuses on a single critical care condition, while the latter relies on static, labeled datasets that do not involve decision-making regarding actions. There is currently no environment capable of providing users with a state where the reward value is derived from the observation in real-world populations.

Aligning LMs and diagnostic reports. Prior studies for medical LMs focus on fitting medical knowledge into model weights. **Alignment by pretraining.** GatorTron trains a clinical LM from scratch on 82 billion words of de-identified notes and improves five clinical NLP tasks Yang et al. (2022), and NYUTron pretrains on 7.25 million notes from a single health system before being fine-tuned per task to predict readmission, mortality, comorbidity, length of stay, and insurance denial, reaching 78.7 to 94.9 AUC and passing a prospective single-arm trial Jiang et al. (2023b). Consequently, solely real outcomes enter LM weights while no RWE reasoning path emerges by pretraining. **Alignment by fine-tuning.** The diagnostic path itself is shaped by instructions written by

clinicians over real EHRs Fleming et al. (2024), by verifier-guided trajectories and RL on 40K verifiable medical problems Chen et al. (2024), by rewards on format and answer matching over medical images Lai et al. (2026); Mullappilly et al. (2026), and by rewards on a ranked differential and consistent examination requests over pathology cases resolved by a knowledge-grounded simulator Schmidgall et al. (2024). While a path is produced, nothing along it is answered by RWE, because the supervision remains an answer key, a clinician ranking, or a simulated examination of a single case rather than observing over real-world populations. There is currently no alignment requiring an LM to reason a diagnostic CoT whose every step is settled by what a matched real-world population experienced. CohortWorld imposes this requirement as an agentic RL environment, where the states are real cohorts and the actions move between them, so the reasoning can be aligned either zero-shot or RL post-training.

3 METHODS

Problem Formulation. Let $\mathcal{P}$ be the finite set of recorded subjects, $\mathcal{F}$ be the finite set of features a record can hold (such as an age range, biological sex, or a code paired with the age range in which it was first recorded), and $\mathcal{D}$ be a fixed set of diseases. For a subject $i \in \mathcal{P}$ and a disease $D \in \mathcal{D}$, let $T_i(D) \in \mathbb{R}^+$ be the time from the date at which the question is asked until subject i is first diagnosed with disease D, taking the value $+\infty$ if never. $\mathcal{H}_i \subseteq \mathcal{F}$ is the set of features recorded before that date, and $\pi_i(D,h) = \Pr[T_i(D) \leq h \mid \mathcal{H}_i] \in [0,1]$ is the disease risk that might be diagnosed within $h \in \{1,2,3,5,10\}$ years in the future. A cohort $c \subseteq \mathcal{P}$ is a group of subjects who all share the same features $\mathcal{H}_c \subseteq \mathcal{F}$, which we call the key of the cohort, and $r_c(D,h) \in [0,1]$ indicates the percentage of members first diagnosed within h years. Hence, the problem in this work is asking LM to answer for $r_c(D,h)$ as a representation of $\pi_i(D,h)$ given $\mathcal{H}_c$.

3.1 GREEDY COHORT COLLECTION

A cohort key $\mathcal{H}_c$ is a 5-axis made of an age range, a sex, and, for each of three code books, a set of pairs of the coded disease or treatment event from the tree-shaped coding system and the age band at which that event first occurred. The code books are the International Classification of Diseases, Tenth Revision (ICD-10) World Health Organization (2004) (2004), the Office of Population Censuses and Surveys Classification of Interventions and Procedures, Fourth Revision (OPCS-4) Connecting for Health (Organization) (2009), and BNF Joint Formulary Committee (2026). Because the index time T_i when the question is asked affects whether a code will appear in the history $\mathcal{H}_i$, we expand the subject pool $\mathcal{P}$ into a pseudo-subject pool using the form $\{(i, T_i)\}$ with the history $\mathcal{H}_{(i,T_i)}$. The candidate index times are from the first appearance of an event in history to every 5-year age gap crossing. We keep a candidate index if it yields one event after T_i.

Cohorts are formed by altering $\mathcal{H}_{(i,T_i)}$ for every pseudo-subject at a chosen level in the coding tree and grouping the pseudo-subjects who share the same $\mathcal{H}_{(i,T_i)}$. The coding tree level is defined as a tuple $\gamma = (\gamma_{\mathrm{I}}, \gamma_{\mathrm{O}}, \gamma_{\mathrm{B}}, \gamma_{\mathrm{A}})$ in which γ_{I}, γ_{O} and γ_{B} each pick a depth of the ICD-10, OPCS-4 and BNF code trees, namely chapter and category for ICD-10, chapter, category and subcategory for OPCS-4, and chapter, section and paragraph for BNF, where γ_{A} picks one of five ways to coarsen the onset age, namely 10-year range, four life stages (by evenly split history into four periods), three life stages (by split into three periods), two life stages (by split into two periods), or no onset information. The $\mathcal{H}_{(i,T_i)}$ at level γ is denoted by $\mathcal{H}^{\gamma}_{(i,T_i)}$. At each γ, the pseudo-subjects not yet placed are grouped by the same, and if one subject i contributes several pseudo-subjects to one group, only the one with the earliest T_i is kept, to prevent disease risk from being inflated by the same subject. A group with at least $k = 10$ members becomes a cohort c whose key $\mathcal{H}_c$ is unique, and its members leave the pool for the collection, whereas the members of a smaller group stay in the pool and are described again at the next coarser γ. Now, we can have exact definition for $\mathcal{H}_c = (b_c, x_c, E^{\mathrm{I}}_c, E^{\mathrm{O}}_c, E^{\mathrm{B}}_c)$, with age $b_c \in \mathcal{B}$, sex $x_c \in \mathcal{X}$ and $E^a_c \subseteq \mathcal{F}^{\gamma_a}_a \times \mathcal{L}_{\gamma_{\mathrm{A}}}$ the pairs of code and life stage recorded on axis a. Besides, each released cohort stores a $5 \times |\mathcal{D}|$ matrix of $r_c(D,h)$ over $h \in \{1,2,3,5,10\}$. The collection of cohorts is used as states in the CohortWorld environment and datasets in CohortCal benchmarking.

3.2 COHORTWORLD RL ENVIRONMENT

State Space. A state is a collected cohort c, and the agent observes the key $\mathcal{H}_c$ serialized as a patient description along with the size of c, which is represented by a range to prevent individuals from being identifiable, the level γ used by c, and the valid action parameters for c, formed as (axis, operation). Let $\mathcal{B}$ denote the set of age bands, $\mathcal{X}$ denote the two sexes, $\mathcal{L}$ denote the life stages a first record can be assigned to, and $\mathcal{F}_a^{\gamma_a}$ denote the codes of 3 coding systems $a \in \{\mathrm{I}, \mathrm{O}, \mathrm{B}\}$. Then,

$$\mathcal{H}_c \in \mathcal{B} \times \mathcal{X} \times \prod_a 2^{\mathcal{F}_a^{\gamma_a} \times \mathcal{L}_{\gamma_\mathrm{A}}}. \tag{1}$$

$r_c(D, h)$ is not observable if that is exactly what the patient agent is diagnosing. Otherwise, the agent can see $r_c(D, h) \in [0, 1]$, along with $\mathcal{N}$, a band of cohort size, and $\mathcal{N}'$, a size band of the members after excluding those who have lost records or are dead.

Action Space. There are four high-level actions in CohortWorld.

`cw_look` `cw_observe {disease} {horizon}`
`cw_move {axis}.{op}.{elem}` `cw_commit {rates}`

where `{axis}` is denoted by $v \in \mathcal{V} = \{\mathrm{AGE}, \mathrm{SEX}, \mathrm{I}, \mathrm{O}, \mathrm{B}\}$, `{elem}` denoted by e, `{disease}` D, `{horizon}` h and `{rates}` $\hat{r}$ are the values those actions read or write, and $\omega \in$ `{op}` is one of six low-level operations:

`add` `switch` `shift` `zoom+` `zoom-` `drop`

i) `cw_look` takes no parameter and reports serialized c with moving parameter candidates. **ii)** `cw_move` moves the agent from one cohort to another. It takes (v, ω, e) as the parameters, resulting in $\tau(\mathcal{H}_c, v, \omega, e)$ to replace $\mathcal{H}_c$. Here `add` inserts e into the set of v, `switch` replaces one element of that set by e, `shift` resets the age band to $e \in \mathcal{B}$, `zoom`$\pm$ moves γ_v one depth, and `drop` deletes e. **iii)** `cw_observe` reads one real-world rate of the cohort the agent stands on, taking $(D, h) \in \mathcal{D} \times \{1, 2, 3, 5, 10\}$ or nothing, in which case it reads the pair the question asks about. **iv)** `cw_commit` states the answer and ends the episode.

Tasks. There are three tasks defined in CohortWorld. The CohortThink task is the diagnosis of Fig. 1b carried out end-to-end, in which a doctor matches a patient's history with a cohort seen in practice, reasons about the differences that remain, and applies their effect on incidence. Its item is a pseudo-subject (i, T_i) serialized from a cohort $c^\star$ together with a disease D and a horizon h, and its ground truth (GT) is $r_{c^\star}(D, h)$, shortened to r in the paper. The agent receives the serialized description and the question, starts on $c^\star$ with $r_{c^\star}$ withheld, and answers in two turns: First the `<cohort-think>` block and then the actions that can be parsed from the thinking block. We read its answer from the first well-formed source among the committed $\hat{r}$, an `ANSWER:` line in its final message, the last rate it read, and a fixed report, and we score $\mathrm{med}(\hat{r})$ against r across items by the two population metrics of Section 3.3.

The CohortMatch task extracts the matching step in the CohortThink item, i.e., whether the agent can find the cohort that matches the patient's history. Its item is the same pseudo-subject, and its truth is the matched cohort. The agent receives the description and an instruction to reach the recorded group whose history differs from the patient's on the fewest features. Detailed extraction steps can be found in Appendix A.2

Table 1: CohortWorld items.

Split	train	seen	unseen
# items	16,770	880	943

The CohortContrast task extracts the effect of a difference in disease risk between a pseudo-subject and a cohort in the CohortThink item. Its item is a pair of cohorts c and c' whose keys differ in exactly one feature, together with D and h, and its truth is the direction of $r_{c'}(D, h) - r_c(D, h)$, labeled lower, higher, or indistinguishable by a Wilson interval on the two rates. The difference prediction is denoted by an unsigned value in $[0, 1]$, where the three directions are cut by 0.25 and 0.75. Table 1 gives the items, assigned to train, seen, and unseen in the ratio $90/5/5$ for CohortThink. The other tasks have the same number of items since they are extracted from CohortThink items.

3.3 CROSS-SAMPLE (CS) GRPO POST-TRAINING

We train the CohortThink policy in two stages: supervised finetuning (SFT) on the doctor-like RWE diagnostic CoT as the cold start of R1 Guo et al. (2025), then reinforcement learning (RL) inside CohortWorld.

Two-turn episode. A model is first SFT-ed on the doctor-like RWE diagnostic CoT to prompt the LM using one `<cohort-think>` block describing the difference between the relevant cohort and the pseudo-subject. Together with the second stage, RL post-training, the two-turn episode is defined as $\tau = (x, y^{(1)}, q, y^{(2)})$. Here x is the serialized question to the LM, $y^{(1)}$ is the first reasoning turn, one block that ends with a signed difference $\delta \in \mathbb{R}$, q is the environment's fixed request to act with CohortWorld tools, and $y^{(2)}$ is the second acting turn. The answer credited to the episode is indicated by $\hat{r} = \max(r_{c'}(D,h) - \delta, 0)$, where c' is the cohort matched by the LM. Therefore, sampled token positions form four sets, $\mathcal{T}^{\phi}$ over $y^{(1)}$, $\mathcal{T}^{\lambda}$ over the tool calls in $y^{(2)}$, $\mathcal{T}^{g}$ over the committed number, and $\mathcal{T}^{\psi} \subset \mathcal{T}^{\phi}$ over δ.

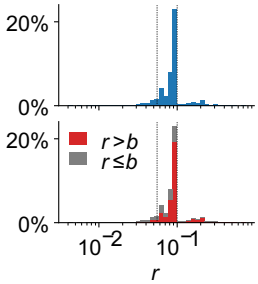

Figure 2: The distribution of GT r.

Population reward. Fig. 2 shows r distribution, where most of $r_c(D,h)$ is small. We therefore score the prediction $\hat{r}$ as a classification task, i.e., whether $\hat{r}$ lies above or below a baseline GT disease risk $b \in (0,1]$ for the same (D,h) over the entire UKB population. As shown in Fig. 2 lower panel, the binary label is more balanced than the r distribution. Each training step samples $M \in \mathbb{N}$ pseudo-subjects from CohortThink items. $\tau_{i,j}$ denotes the j-th response of the agent with $\hat{r}_{i,j}$ the predicted disease risk, $i \in \{1,\dots,M\}$ and $j \in \{1,\dots,N\}$. Along j, the N responses to one pseudo-subject are the group of the original GRPO Shao et al. (2024), and the format and invocation rewards of the `<cohort-think>` block are credited within it. Along i, the M responses to different pseudo-subjects, $\hat{r}_{1,j},\dots,\hat{r}_{M,j}$, are a population, and the AUROC and calibration slope are computed on it and credited to each $\tau_{i,j}$ by its own contribution, $g_{i,j}$. A population's AUROC is piecewise constant in any one of its $\hat{r}$ Cortes & Mohri (2003), so a reward equal to it would be set by the other members rather than by $\tau_{i,j}$. We use $g_{i,j}$ to denote the contribution of $\hat{r}_{i,j}$ to the population's AUROC and slope score. With $s_{i,j} = \hat{r}_{i,j}/b_i$, P and Q the members of the population with $r > b$ and with $r \le b$,

$$g_{i,j} = \tfrac{1}{2}\big(a_{i,j} + \kappa_{i,j}\big), a_{i,j} = \begin{cases} \dfrac{1}{|Q|}\displaystyle\sum_{o\in Q}\Big(\mathbf{1}[s_{i,j} > s_{o,j}] + \tfrac{1}{2}\,\mathbf{1}[s_{i,j} = s_{o,j}]\Big) & i \in P, \\ \dfrac{1}{|P|}\displaystyle\sum_{o\in P}\Big(\mathbf{1}[s_{o,j} > s_{i,j}] + \tfrac{1}{2}\,\mathbf{1}[s_{o,j} = s_{i,j}]\Big) & i \in Q, \end{cases} \tag{2}$$

where $a_{i,j} \in [0,1]$ is the percentage of the opposite-label members that the pseudo-subject is ordered correctly against, and $\mathrm{AUROC} = \frac{1}{|P|}\sum_{i\in P} a_{i,j} = \frac{1}{|Q|}\sum_{i\in Q} a_{i,j}$ by the Mann-Whitney identity Mann & Whitney (1947), $\kappa_{i,j} = e^{-|\log \hat{\beta}|}$ is the slope score with $\hat{\beta} \in \mathbb{R}$, which is the correlation between ranking timed by standard deviation ratio of the population. Together with the two rewards credited along j, the reward of $\tau_{i,j}$ is

$$R_{i,j} = w_\phi\,\phi_{i,j} + \lambda_{i,j} + w_g\,g_{i,j} + w_\psi\,\psi_{i,j}, \tag{3}$$

where $\phi_{i,j} \in [0,1]$ is a format reward following Guo et al. (2025) containing five requirements on $r^{(1)}$: one block, a closing line stating δ to four decimals, at least twenty words of reasoning before it, no other digit in the reasoning, and no code or jargon. $\lambda_{i,j} \ge 0$ is an invocation reward as in Qian et al. (2026), a fixed amount for issuing any tool call, more for reaching a cohort $c \neq c^{\star}$, and more again for reading $r_c(D,h)$ there. The last term denotes the absolute error of the predicted disease risk difference,

$$\psi_{i,j} = \exp\Big(-\big|\,\delta_{i,j} - \big(r_{c'}(D,h) - r\big)\big| \,/\, u\Big) \in (0,1], \tag{4}$$

with $u > 0$ a fixed scale, and $\psi_{i,j} = 0$. It scores the second turn in CohortThink with the feedback from the real-world disease risk difference, and $w_\phi, w_g, w_\psi > 0$ are fixed hyperparameters (see Appendix A.4).

Cross-sample advantage and objective. The advantage $A_{i,j,t}$ weights the log-probability gradient of token t in the policy update. Each part of equation 3 becomes one by centering along j-axis as in

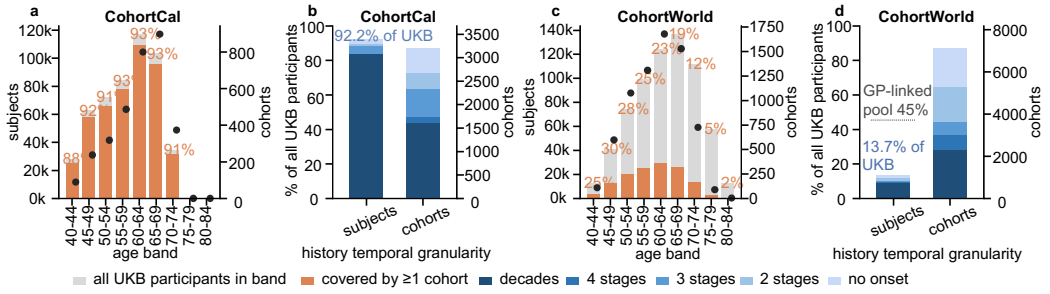

Figure 3: Subjects and released cohorts in CohortWorld and CohortCal.

GRPO Shao et al. (2024) without the standard deviation normalization Chen et al. (2026), and lands on its own span, $A_{i,j,t} = w_\phi(\phi_{i,j} - \bar{\phi}_i)\mathbf{1}[t \in \mathcal{T}^{\phi}_{i,j}] + (\lambda_{i,j} - \bar{\lambda}_i)\mathbf{1}[t \in \mathcal{T}^{\lambda}_{i,j}] + w_g(g_{i,j} - \bar{g}_i)\mathbf{1}[t \in \mathcal{T}^{g}_{i,j}] + w_\psi(\psi_{i,j} - \frac{1}{2})\mathbf{1}[t \in \mathcal{T}^{\psi}_{i,j}]$, where t indexes token positions and a bar is the mean over the N responses to one pseudo-subject. $g_{i,j}$ is thus computed along i-axis and centered along j-axis, which wraps up as the cross-sample step. $A_{i,j,t}$ enters the importance-weighted policy gradient unchanged, with the cross-entropy on the tokens of the tool results Intellect (2025).

4 Results

CohortCal and CohortWorld datasets. We built CohortCal as a public benchmark to evaluate the alignment between RWE diagnostics and LMs by reasoning on medical history. It contains 70,378 items from total 3,198 cohorts sampled from all collected cohorts (see Appendix A.3). Table 2 lists the split details.

Table 2: CohortCal items.

Split	train	seen	unseen
# items	32,120	6,474	31,784

Fig. 3 shows that CohortCal is a population-scale benchmark rather than a narrow subset of UKB. Its 3,198 cohorts cover 462,802 of 502,000 UKB subjects, or 92.2%, where $\mathcal{H}_c$ can miss OPCS-4 or BNF since some subjects have no GP linkage. In Fig. 3a, coverage remains high across the adult age bands with sufficient follow-up, while the black markers show that the benchmark retains many distinct cohorts within each band. Fig. 3b further shows that most subjects retain decade-level γ_A, resulting in CohortCal evaluating diagnostic risk across both detailed and deliberately generalized histories. On the other hand, the CohortWorld state space requires no axis to be missing. Fig. 3c and d show 7,113 cohorts from the 227,727 UKB subjects with GP linkage and histories in all three coding systems. It covers 68,577 subjects, or 13.7% of UKB. Coverage concentrates in ages with linked longitudinal records and falls at the oldest ages, while the cohorts span all five γ_A. This restricted cohort release enables more trustworthiness as an actionable world. CohortCal cohorts do not overlap with CohortWorld cohorts, so there is no data leakage by using the CohortWorld action for the CohortCal benchmark. The RL environment with all datasets is available at `https://anonymous.4open.science/r/cohort-world-E44D`.

Table 3: CohortCal benchmarking. Metrics are from unseen split, where we report 95% bootstrap intervals.

model	AUROC	beats b	slope
Qwen3.5-2B	$0.490_{[0.483,0.497]}$	0.319	0.046
Qwen3.5-4B	$0.484_{[0.477,0.491]}$	0.333	0.053
Qwen3.5-9B	$0.543_{[0.537,0.550]}$	0.352	0.206
Qwen3.6-27B	$0.514_{[0.507,0.521]}$	0.277	0.178
Gemma 3 4B	$0.441_{[0.434,0.448]}$	0.335	0.010
Gemma 3 27B	$0.517_{[0.510,0.524]}$	0.328	0.129
MedGemma 4B	$0.499_{[0.493,0.507]}$	0.330	0.010
MedGemma 27B	$0.556_{[0.550,0.563]}$	0.328	0.197
Llama 3 8B	$0.534_{[0.528,0.541]}$	0.359	0.104
Mistral Small 24B	$0.544_{[0.537,0.551]}$	0.370	0.130
DeepSeek-V4.1-Flash	$0.567_{[0.561,0.574]}$	0.373	0.211
Claude Sonnet 5	$0.560_{[0.554,0.567]}$	0.338	0.345
Gemini 3.5 Flash-Lite	$0.507_{[0.500,0.514]}$	0.240	**0.536**
Qwen3.5-9B w/ CW	$0.618_{[0.611,0.625]}$	0.461	0.389
Qwen3.6-27B w/ CW	$\mathbf{0.668}_{[\mathbf{0.662,0.675}]}$	**0.504**	0.407

Language Models. On CohortCal, we evaluate 11 open-weight models from five families, Qwen3.5 Team (2026), Gemma 3 Team et al. (2025), MedGemma Sellergren et al. (2025), Llama 3 Repede & Brad (2024), Mistral Small Jiang et al. (2023a) and DeepSeek DeepSeek-AI (2026), and two closed-source models, Claude Sonnet 5 with low-level effort and Gemini 3.5 Flash-Lite with minimal thinking mode to make the leaderboard. Each of them answers in a single turn from the serialized history as shown in the example questions in Fig 5 and predicts $\hat{r}$. One open-weight model,

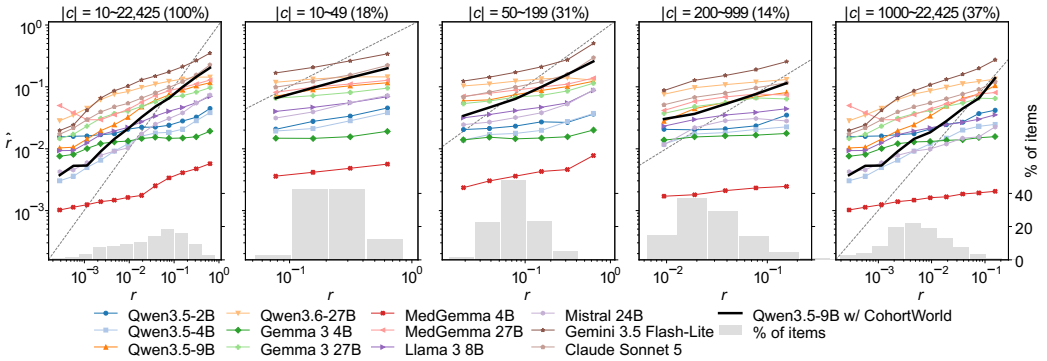

Figure 4: Predicted $\hat{r}$ against ground truth r across cohort-size bands on CohortCal benchmark. Grey bars show the percentage of items in each bin.

Qwen3.5-9B, is tested using CohortWorld actions for CohortCal. On the CohortWorld tasks, we take two open-weight models to show the generalizability, Qwen3.5-9B and Qwen3.6-27B, and run each in three ways: (i) the same single-turn as CohortCal, (ii) a CohortWorld agent without post-training, and (iii) with CS-GRPO post-training. Appendix A.5 and A.4 give the prompts and the settings for both CohortCal and CohortWorld.

4.1 LEADERBOARDS

Table 3 lists the leaderboard of CohortCal. AUROC indicates the classification performance of whether $\hat{r} > b$ or $\hat{r} \leq b$ for the baseline disease risk b in the broader population, regardless of the medical history difference. Appendix A.6 defines metrics in detail. Claude Sonnet 5 was the strongest single-turn model by AUROC at 0.560. The Qwen3.6-27B with CohortWorld (CW) equipped reached the best AUROC, 0.668. The usage of CohortWorld also increases 'beats b', denoting the percentage of diagnoses by LM closer to r than a population average b, from 0.338 to 0.504, and reaches a second-place slope of 0.407, which is the coefficient of $\hat{r}$ and r.

Table 4: CohortWorld tasks benchmarking.

model	CM acc.	CC acc.	CT AUROC	Tool calls
Qwen3.5-9B	6.9%	**59.8%**	$0.427_{[0.390,0.464]}$	0
w/ CW	29.6%	51.6%	$0.516_{[0.479,0.553]}$	$12.0_{\pm 7.4}$
+ CS-GRPO	80.2%	16.0%	$0.621_{[0.591,0.655]}$	$3.3_{\pm 1.6}$
+ ψ	**80.0%**	16.3%	$\mathbf{0.638}_{[\mathbf{0.605},\mathbf{0.670}]}$	$3.3_{\pm 1.9}$
Qwen3.6-27B	7.1%	**67.2%**	$0.450_{[0.413,0.486]}$	0
w/ CW	22.0%	61.1%	$0.621_{[0.582,0.656]}$	$9.7_{\pm 6.0}$
+ CS-GRPO	77.0%	20.5%	$\mathbf{0.642}_{[\mathbf{0.609},\mathbf{0.674}]}$	$3.4_{\pm 2.3}$
+ ψ	**77.4%**	18.1%	$0.627_{[0.594,0.660]}$	$3.2_{\pm 1.2}$

Figure 4 groups CohortCal items by cohort size and plots mean $\hat{r}$ in log-spaced bins of r, with the diagonal at $\hat{r} = r$ and grey bars for the percentage of items in each bin. Every single-turn model is flat over all groups of r. They all sit above the diagonal at the low rates, and therefore report more risk than the ground truth. As $|c|$, the cohort size, increases, they move from below the diagonal in the smallest band to above it in the largest. The CohortWorld curve is the nearest to the diagonal in most of cases.

Table 4 evaluates CohortMatch (CM), CohortContrast (CC), and CohortThink (CT) on the unseen split. CM accuracy (acc.) indicates the exact matching rate as defined in Section 3.2 for reasoning. CC acc. indicates the three-way direction classification. CT is the same question as CohortCal, reporting AUROC. CT examples are shown in Appendix A.2, where the extraction of CC and CM are also demonstrated. Without CW, Qwen models achieve 6.9% and 7.1% on CM acc. for both model sizes. In contrast, the zero-shot agents reach 29.6% and 22.0% CM acc., and CS-GRPO takes Qwen3.5-9B to 80.2%. CC moves the other way, e.g., from 51.6% to 16.0%. This indicates that CS-GRPO buys the matching step and gives up the contrast step, implying that cohort matching affects RWE diagnostics more than risk difference estimation. On CT, the two single-turn rows sit below chance at 0.427 and 0.450, CohortWorld alone lifts them to 0.516 and 0.621, and CS-GRPO raises them further to 0.638 and 0.642. The gain from training is 0.12 at 9B and 0.02 at 27B, where the zero-shot agent already performs good cohort matching, so CS-GRPO, which focuses on the matching step, mainly boosts the small Qwen. The number of tool calls indicates the cost of CW.

CohortThink Examples. Figure 5 shows the three example outcomes from Claude Sonnet 5 and Qwen3.5-9B on two items, and more detailed examples can be found in Appendix A.2. The ground truths of the two items differ by more than 200%, 0.027 and 0.067, while Claude Sonnet 5 answers

Legends | question | <cohort-think> reasoning | **tool call {arguments}** | wrong or unsupported | answer

Claude Sonnet 5	Qwen3.5-9B in CohortWorld	Qwen3.5-9B in CohortWorld after CS-GRPO
Q. Consider UK adults aged 60–64 (male; hospital records only) with the following recorded medical history (conditions grouped by the age at which they first appeared): - 50–60: hypertension. **To Claude Sonnet 5:** Among such people who have NOT previously been diagnosed with this condition, what percentage are diagnosed with "musculoskeletal disorder [M0]" (ICD-10 M0) for the FIRST time within the next 10 years? Answer with a single percentage. **In CohortWorld:** Your task is to: estimate the probability that he is diagnosed with "musculoskeletal disorder" for the first time in the next 10 years. He has never been diagnosed with it. Reason about which larger group of patients you would compare this one against, and by how much this patient's own risk differs from that group's.		
A. 4.5	**A.** <cohort-think> The user wants to estimate the probability … I need to find where I am currently "standing" in the lattice …</cohort-think> **cw_look {}** … Hypertension is typically coded as ICD10 code I10 … [2,000 words, ending in repeated text] … actual prompt text clearly states "next 1 year" **cw_move {bnf.add.0206}** **cw_observe {disease: "musculoskeletal disorder", horizon: "1"}** → error, no such disease **cw_observe {disease: "M", horizon: "1"}** → error, budget exhausted → fixed report 0.50	**A.** <cohort-think>Compared with people who share this history and in addition have medication for cardiovascular, this patient's risk is higher. The presence of hypertension suggests a higher likelihood of underlying vascular issues … Adjustment: +0.0008.</cohort-think> **cw_look {}** → **cw_move {bnf.add.0205}** (antihypertensives, 50–60) → **cw_observe {}** → NO_FOLLOWUP_WINDOW → **cw_move {bnf.zoom_coarser}** (cardiovascular [02], n 50–199) → **cw_observe {}** → rate 0.0270, at risk 30–49 → **cw_commit {}**
Scored 0.045, **GT** 0.027	**Scored** 0.50 (fixed report), **GT** 0.027	**Scored** 0.027, **GT** 0.027
Q. Consider UK adults aged 65–69 (male; hospital records only) with the following recorded medical history (conditions grouped by the age at which they first appeared): - 60+: atrial fibrillation, hypertension. **To Claude Sonnet 5:** Among such people who have NOT previously been diagnosed with this condition, what percentage are diagnosed with "infectious disease [A4]" (ICD-10 A4) for the FIRST time within the next 10 years? Answer with a single percentage. **In CohortWorld:** Your task is to: estimate the probability that he is diagnosed with "infectious disease" for the first time in the next 10 years. He has never been diagnosed with it. Reason about which larger group of patients you would compare this one against, and by how much this patient's own risk differs from that group's.		
A. 4.2	**A.** (no block) **cw_look {}** … Common BNF codes for cardiovascular conditions include codes in the 0103 range … **cw_move {bnf.add.0103}** → "gastro-intestinal [0103]" … **cw_move {bnf.switch.0212}** → cardiovascular [0212] … **cw_observe {disease: "infectious disease", horizon: "10"}** → rate 0.0, at risk 22–29 … **cw_move {bnf.add.0501}** → "infections [0501]" … This is not what I want …	**A.** <cohort-think>Compared with people who share the rest of this history but for whom atrial fibrillation is not required, this patient's risk is higher. The presence of atrial fibrillation often indicates underlying heart issues or systemic inflammation, which can impair immune function … Adjustment: -0.0009.</cohort-think> **cw_look {}** → **cw_move {icd.add.I48}** (atrial fibrillation, 60+) → **cw_observe {}** → NO_FOLLOWUP_WINDOW → **cw_move {icd.zoom_coarser}** (circulatory [I], n 10–49) → **cw_observe {}** → rate 0.0667, at risk 13–16 → **cw_commit {}**
Scored 0.042, **GT** 0.067	**Scored** 0.0, **GT** 0.067	**Scored** 0.067, **GT** 0.067

Figure 5: Two examples of CohortCal item with question-answering showcases.

4.5% and 4.2%, two closing numbers. Prevalence predictions derived from recalled medical knowledge in Sonnet 5 do not align with RWE in UKB. Qwen3.5-9B in CohortWorld without CS-GRPO post-training observed no follow-up window because it inappropriately matched to a small cohort or misused the CohortWorld action. After post-training, the agent takes proper diagnostic reasoning before proceeding with actions and successfully matched to relevant cohorts, leading to an accurate disease risk diagnosis. This reflects the high importance of the CohortMatch task for the final CohortThink outcome, agreeing with the quantitative results in Table 4.

5 Conclusions and Discussion

We introduced CohortWorld, an environment for aligning LMs with real-world evidence (RWE) in disease-risk diagnostics. Existing medical benchmarks and virtual environments often rely on synthetic data, answer keys, or inaccessible patient records. CohortWorld instead represents longitudinal health data as privacy-preserving cohorts. It lets agents match medical histories, retrieve observed cohort-level risks, and estimate disease risk through explicit actions. We also introduced CohortCal, a benchmark for evaluating whether language models produce calibrated disease-risk estimates from medical histories, and CS-GRPO, a post-training method that rewards agents for grounding their RWE diagnostic estimates in retrieved cohort evidence. Our experiments show that general and medical LMs have weak risk calibration when they rely on medical histories alone. CohortWorld improves discrimination and calibration. Post-training further improves performance on cohort matching. However, several limitations remain. The environment is based solely on UK Biobank and may not generalize to other populations. The current histories omit biomarkers. Together, CohortWorld provides an evidence-based benchmark for quantifying the hallucination by LMs on diagnostics with accessible datasets and an RL environment.

AI USE STATEMENT

In this work, we used generative AI tools for coding and writing polishing. We have not used generative AI tools for figure generation, dataset construction, or ideation. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness by the first author. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses UK Biobank under an approved application (# is hidden for double blind review), and no individual-level record is released. A released cohort is a cohort of at least $k = 10$ participants described by an age band, a sex, and coded history at a chosen depth of the coding trees. Cohort sizes are published as ranges, and the released quantity is an incidence over the cohort. A pseudo-subject is a history and an index time drawn from that construction and not a person's record, so the examples in this paper are cohort descriptions and not patient notes.

REPRODUCIBILITY STATEMENT

Section 3.1 defines the cohort collection. Section 3.2 defines the environment, its four actions, the three tasks and their scoring, and Section 3.3 defines the reward and the advantage of CS-GRPO. Appendix A.1 gives the templates and the data generation of the cold-start SFT data, Appendix A.5 the prompts and settings of every model, and Appendix A.4 every setting of the cold start and of both reinforcement learning runs, including all hyperparameters. Appendix A.3 details the cohort counts of Section 4, and Appendix A.6 defines every metric. We release the environment, the cohort collection, the benchmark items, and the evaluation harness at the address in Section 4. A recorded seed fixes the splits, and no cohort key appears in two of them, so a rerun scores the same items. Every interval in this paper is a percentile bootstrap with $2{,}000$ resamples at seed 42.

ACKNOWLEDGMENTS

Funding information is hidden for double blind review.

REFERENCES

Carla AbouZahr, Lene Mikkelsen, Rasika Rampatige, and Alan Lopez. Mortality statistics: a tool to improve understanding and quality. *Health information systems*, 2010.

Yan Cai, Linlin Wang, Ye Wang, Gerard De Melo, Ya Zhang, Yanfeng Wang, and Liang He. Medbench: A large-scale chinese benchmark for evaluating medical large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17709–17717, 2024.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*, 2024.

Xiwen Chen, Wenhui Zhu, Peijie Qiu, Xuanzhao Dong, Hao Wang, Haiyu Wu, Huayu Li, Aris Sotiras, Yalin Wang, and Abolfazl Razi. Dra-grpo: Your grpo needs to know diverse reasoning paths for mathematical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 13995–14019, 2026.

Connecting for Health (Organization). *Opcs Classification of Interventions and Procedures Version 4.5 (April 2009): Alphabetical V. 2*. OPCS Classification of Interventions and Procedures, Version 4.5. TSO, 2009. ISBN 9780113228317. URL https://books.google.com/books?id=7MO1wJJHO1wC.

Corinna Cortes and Mehryar Mohri. Auc optimization vs. error rate minimization. *Advances in neural information processing systems*, 16, 2003.

DeepSeek-AI. Deepseek-v4.1-flash: Pushing the limits of kv cache compression, 2026.

Scott L Fleming, Alejandro Lozano, William J Haberkorn, Jenelle A Jindal, Eduardo Reis, Rahul Thapa, Louis Blankemeier, Julian Z Genkins, Ethan Steinberg, Ashwin Nayak, et al. Medalign: A clinician-generated dataset for instruction following with electronic medical records. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 22021–22030, 2024.

Jessica M Franklin, Elisabetta Patorno, Rishi J Desai, Robert J Glynn, David Martin, Kenneth Quinto, Ajinkya Pawar, Lily G Bessette, Hemin Lee, Elizabeth M Garry, et al. Emulating randomized clinical trials with nonrandomized real-world evidence studies: first results from the rct duplicate initiative. *Circulation*, 143(10):1002–1013, 2021.

Omer Gottesman, Fredrik Johansson, Joshua Meier, Jack Dent, Donghun Lee, Srivatsan Srinivasan, Linying Zhang, Yi Ding, David Wihl, Xuefeng Peng, et al. Evaluating reinforcement learning algorithms in observational health settings. *arXiv preprint arXiv:1805.12298*, 2018.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Mason Hargrave, Alex Spaeth, and Logan Grosenick. Epicare: A reinforcement learning benchmark for dynamic treatment regimes. *Advances in neural information processing systems*, 37:130536–130568, 2024.

Prime Intellect. prime-rl, 2025. URL https://github.com/PrimeIntellect-ai/prime-rl.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023a. URL https://arxiv.org/abs/2310.06825.

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

Yixing Jiang, Kameron C Black, Gloria Geng, Danny Park, James Zou, Andrew Y Ng, and Jonathan H Chen. Medagentbench: a virtual ehr environment to benchmark medical llm agents. *Nejm Ai*, 2(9):AIdbp2500144, 2025.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. Swe-bench: Can language models resolve real-world github issues? In *International Conference on Learning Representations*, volume 2024, pp. 54107–54157, 2024.

Joint Formulary Committee. *British National Formulary (BNF 91) March 2026*. Pharmaceutical Press, 2026. ISBN 9780857114969. URL https://books.google.com/books?id=63Xk0QEACAAJ.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Matthieu Komorowski, Leo A Celi, Omar Badawi, Anthony C Gordon, and A Aldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature medicine*, 24(11):1716–1720, 2018.

Nicholas I-Hsien Kuo, Mark N Polizzotto, Simon Finfer, Federico Garcia, Anders Sönnerborg, Maurizio Zazzi, Michael Böhm, Rolf Kaiser, Louisa Jorm, and Sebastiano Barbieri. The health gym: synthetic health-related datasets for the development of reinforcement learning algorithms. *Scientific data*, 9(1):693, 2022.

Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, Yuheng Li, Konstantinos Psounis, and Xiaofeng Yang. Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. *IEEE transactions on medical imaging*, 2026.

Zhiyao Luo, Mingcheng Zhu, Fenglin Liu, Jiali Li, Yangchen Pan, Jiandong Zhou, and Tingting Zhu. Dtr-bench: An in silico environment and benchmark platform for reinforcement learning based dynamic treatment regime. *arXiv preprint arXiv:2405.18610*, 2024.

Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pp. 50–60, 1947.

Sahal Shaji Mullappilly, Mohammed Irfan Kurpath, Omair Mohamed, Mohamed Zidan, Fahad Khan, Salman Khan, Rao Anwer, and Hisham Cholakkal. Medix-r1: Open ended medical reinforcement learning. *arXiv preprint arXiv:2602.23363*, 2026.

Cheng Qian, Emre Can Acikgoz, Qi He, Hongru Wang, Xiusi Chen, Dilek Hakkani-Tur, Gokhan Tur, and Heng Ji. Toolrl: Reward is all tool learning needs. *Advances in Neural Information Processing Systems*, 38:105523–105553, 2026.

Stefan Emil Repede and Remus Brad. Llama 3 vs. state-of-the-art large language models: Performance in detecting nuanced fake news. *Computers*, 13(11):292, 2024.

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

Beth Schurman. The framework for fda's real-world evidence program. *Appl Clin Trials*, 28(4): 15–17, 2019.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Karan Singhal, Tao Tu, Juraj Gottweisguo2025deepseek, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature medicine*, 31(3):943–950, 2025.

Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779, 2015.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Qwen Team. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

Trieu H Trinh, Yuhuai Wu, Quoc V Le, He He, and Thang Luong. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482, 2024.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023.

Xumeng Wen, Zihan Liu, Shun Zheng, Shengyu Ye, Zhirong Wu, Yang Wang, Zhijian Xu, Xiao Liang, Junjie Li, Ziming Miao, et al. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms. In *International Conference on Learning Representations*, volume 2026, pp. 49450–49483, 2026.

World Health Organization (2004). *ICD-10: international statistical classification of diseases and related health problems: tenth revision*. World Health Organization, 2004.

Michael Wornow, Rahul Thapa, Ethan Steinberg, Jason Fries, and Nigam Shah. Ehrshot: An ehr benchmark for few-shot evaluation of foundation models. *Advances in Neural Information Processing Systems*, 36:67125–67137, 2023.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh J Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, et al. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Advances in Neural Information Processing Systems*, 37:52040–52094, 2024.

Hang Yang, Hao Chen, Hui Guo, Yineng Chen, Ching-Sheng Lin, Shu Hu, Jinrong Hu, Xi Wu, and Xin Wang. Llm-medqa: Enhancing medical question answering through case studies in large language models. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2025.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B Costa, Mona G Flores, et al. A large language model for electronic health records. *NPJ digital medicine*, 5(1):194, 2022.

A Appendix

A.1 SFT Data Preparation

The cold start is trained on 1,753 demonstration data with a further 92 held out, drawn from all three depths of the collection, 680 at γ=top level, 552 at γ=1st-depth level and 521 at the γ=2nd-depth level.

The cohort pair in the matching step is chosen by either widening to the largest cohort whose rate is published or dropping the history feature least related to the disease that is asked by using MedGemma-27B. It is avoided that cohort's difference happens to be closest to the truth, since that would teach a selection rule no policy could execute.

The first sentence and the last are hard-coded. The writer is given the facts and returns only the sentences in between, and the demonstration is assembled as

```
opening + writer_text + " Adjustment: {delta:+.4f}."

opening = "Compared with {cohort}, this patient's risk of a first
           diagnosis of {disease} within {h} year(s) is {word}."
word    in {lower, higher, much the same}
```

The cohort is always one valid **cw_move** from the patient's own cohort, and it is named by the operation that reaches it, in the environment's own words.

```
drop     "people who share the rest of this history but for whom
          {feature} is not required"
add      "people who share this history and in addition have {feature}"
switch   "people who share the rest of this history but who have
          {feature} in its place"
shift    "people with the same recorded history who are in the
          {band} age group"
```

where a feature is one entry of the cohort key on one of the three coding axes, written in words rather than as a code, so it reads as a prescription such as medication for cardiovascular, a diagnosis such as musculoskeletal disorder, or a procedure. Four of the six operations of **cw_move** occur: 576 demonstrations by **drop**, 546 by **add**, 475 by **switch** and 156 by **shift**. Because the direction word and the closing δ are both computed from the two cohorts, a demonstration states the comparison and its size as they are, whatever MedGemma-27B argues. A positive δ means the patient's risk is below the rate the agent will read, because the credited answer is $\max(r_{c'}(D,h) - \delta, 0)$. Across all 1,753 demonstrations, this stated δ equals the recorded difference between the two cohorts exactly, which is what makes the comparison in a demonstration correct by construction rather than by a writer's judgement.

MedGemma-27B generates the sentences between the first and the last, reasoning based on its medical knowledge about the relationship between the difference between two cohorts and the target disease. A verbal statement of the difference effect size is also included. MedGemma-27B does not see the GT r. Three examples of SFT demonstration data are shown as below: one where the patient's risk falls below the cohort reached and two where it rises above it.

<cohort-think>Compared with people who share the rest of this history but for whom medication for infections is not required, this patient's risk of a first diagnosis of nervous system within 5 years is lower. This comparison is appropriate because the history of infection treatment suggests a chronic inflammatory process, which is known to be associated with a reduced likelihood of developing certain neurological conditions. The ongoing management of infections likely modulates the immune system in a way that protects against these specific types of nervous system problems. Therefore, the difference in risk is relatively small, perhaps a few cases per thousand people over five years, reflecting the protective effect of controlling the infection. Adjustment: +0.0032.</cohort-think>

<cohort-think>Compared with people who share the rest of this history but for whom medication for gastro-intestinal is not required, this patient's risk of a first diagnosis of digestive within 1 year is higher. This comparison is appropriate because the patient's only recorded history involves medication for gastro-intestinal issues, suggesting an underlying, perhaps symptomatic, condition. Taking medication for digestive problems often indicates that there is an existing issue, even if it hasn't yet led to a formal diagnosis, and this underlying problem increases the likelihood of a new digestive diagnosis being made. The difference in risk is quite substantial, suggesting the medication use is a strong indicator of an increased susceptibility to developing further digestive conditions. Adjustment: -0.0808.</cohort-think>

<cohort-think>Compared with people who share the rest of this history but for whom medication for ENT is not required, this patient's risk of a first diagnosis of musculoskeletal within 5 years is higher. This comparison is appropriate because the patient's only recorded history involves prescriptions related to ear, nose, and throat issues, suggesting a potential underlying inflammatory or autoimmune process. Chronic inflammation, often associated with conditions requiring ENT medication, can increase the likelihood of developing musculoskeletal problems over time. The difference in risk between this patient and that group is notable, suggesting the ENT medication use is a significant factor, perhaps nearly tripling the chance of a new musculoskeletal diagnosis in the coming years. Adjustment: -0.0322.</cohort-think>

A.2 COHORTWORLD EXAMPLES

CohortMatch (CM) and CohortContrast (CC) are extracted from the same CohortThink items. Each item in CM and CC carries a reference cohort, which is the target cohort denoted by r^{ref} and recognized in the same way as SFT. An answer $\hat{r}$ scores CM when it equals r^{ref}, and it scores CC when $\hat{r} - r^{\text{ref}}$ has the sign of $r - r^{\text{ref}}$, with items at $r = r^{\text{ref}}$ excluded from CC. Two items are shown below. Gray text is what the model receives: the question, the follow-up turn of the two-turn protocol, and the results returned by the tools. Every other color is the model's own output, its reasoning and tool calls exactly as generated, one white card per Qwen3.5-9B row of Table 4, under the gray box of the question. Tool results are shortened to the fields the model acted on, and a run that exhausted its budget is cut with [...]. The one-turn prompt asks for a percentage, so 1.8 is the answer 0.018.

Consider UK adults about 77 age (male; linked primary-care and hospital records) with the following recorded history:
- diagnoses: digestive [K] (first recorded 60+)
- procedures: soft tissue [T] (first recorded 60+)
- prescriptions: none recorded

Your task is to: estimate the probability that he is diagnosed with "endocrine/metabolic" for the first time in the next 5 years. He has never been diagnosed with it.

Qwen3.5-9B, no tools

1.8

Qwen3.5-9B w/ CW

cw_look()
cw_look → {"state_id": "C1:3ebde0b12f648c5f", "age_band": "75-79", "sex": "Male", "icd": [K], "opcs": [T], "n_band": "10-49", "legal_moves": {"opcs.drop": [{"id": "opcs.drop.T->C1:e5f7a29d90c8df37", "size_after": "200-999"}], "opcs.switch": ..., "icd.drop": ..., "age_band.shift": ...}}
I am currently in a cohort of UK adults aged 75-79, male, with digestive diagnoses (K) and soft tissue procedures (T) recorded. I need to estimate the risk of a first-time diagnosis of "endocrine/metabolic" in the next 5 years. [...] Let me try dropping one of these conditions first and see what new options arise. Dropping icd.K (digestive) first seems reasonable as it's one of the defining features:
cw_move("move_id": "icd.drop.K->C1:e5f7a29d90c8df37")
cw_move → {"hit": true, "cost": 3.0, "state_id": "C1:e5f7a29d90c8df37", "icd": [], "opcs": [], "bnf": [], "n_band": "200-999", "diseases_here": 19, "moves_left": 24}
Now I'm in a cohort of UK adults aged 75-79, male, with no recorded diagnoses, procedures, or prescriptions. [...] Let me try to observe the rate for endocrine/metabolic disease at horizon 5 (since

we're interested in the next 5 years):
cw_observe("disease": "endocrine/metabolic", "horizon": "5")
cw_observe → {"status": "published", "disease": "E", "disease_name": "endocrine/metabolic [E]", "horizon": 5, "rate": 0.198864, "at_risk": "200-999"}
Excellent! I found the rate for endocrine/metabolic disease [E] in this cohort (UK adults aged 75-79, male, with no recorded diagnoses/procedures/prescriptions) at horizon 5 years. The rate is approximately 0.198864, or about 19.89%. [...] Digestive diseases (K) could be related to metabolic conditions like diabetes (which is endocrine/metabolic). [...] However, the patient's digestive diagnosis was first recorded at age 60+, which is before the age we're considering (77). This suggests the digestive condition has been present for a while. [...] I'll estimate that my patient's risk is somewhat higher than the baseline due to potential overlap between digestive and metabolic conditions:
cw_commit("rates": "[0.15, 0.18, 0.22, 0.25, 0.28]")

Qwen3.5-9B w/ CW + CS-GRPO

<cohort-think>Compared with people who share this history and in addition have medication for cardiovascular, this patient's risk of a first diagnosis of endocrine/metabolic within 5 years is lower. This comparison group is appropriate because their underlying digestive issues and soft tissue procedures suggest similar potential for systemic inflammation or metabolic disturbances that might manifest as endocrine or metabolic problems. The presence of cardiovascular medication in that group points towards a higher likelihood of conditions like diabetes or thyroid issues, which are common comorbidities with cardiovascular disease. Therefore, the absence of this medication in our patient suggests a substantially lower chance of developing an endocrine or metabolic disorder, perhaps by several hundred cases per thousand people. Adjustment: +0.0004.</cohort-think>
You have said which group you would compare this patient against and by how much their risk differs. Now carry it out. The moves have ids you cannot guess, so look first, and reply with nothing but this function call: [...] The result will list every move this group accepts. Then take the one that reaches the group you named, read the rate there with cw_observe, and call cw_commit with that reading minus the difference you stated.
cw_look()
cw_look → {"state_id": "C1:3ebde0b12f648c5f", "icd": [K], "opcs": [T], "legal_moves": {"opcs.drop": [{"id": "opcs.drop.T->C1:e5f7a29d90c8df37", ...}], ...}}
cw_move("move_id": "opcs.drop.T->C1:e5f7a29d90c8df37")
cw_move → {"hit": true, "cost": 2.0, "state_id": "C1:e5f7a29d90c8df37", "icd": [], "opcs": [], "bnf": [], "n_band": "200-999"}
cw_observe()
cw_observe → {"status": "published", "disease": "E", "disease_name": "endocrine/metabolic [E]", "horizon": 5, "rate": 0.198864, "at_risk": "200-999"}
cw_commit()

Extraction. The recorded answer is $r = 0.2667$. True CM is the reference rate $r^{\text{ref}} = 0.1989$ and true CC is the direction $\text{sign}(r - r^{\text{ref}}) = +$. No tools: CM extracted 0.0180, miss; CC extracted $-$, wrong. With CohortWorld: CM extracted 0.22, miss; CC extracted $+$, correct. After CS-GRPO: CM extracted 0.1989, exact; CC extracted 0, wrong. The post-trained model reports the reference rate exactly, scoring CM, and by the same fact carries no direction, missing CC. Both agents read the same cohort, which the base agent reaches by dropping the diagnosis and the post-trained one by dropping the procedure, since either drop empties the history at this depth.

Consider UK adults about 52 age (male; linked primary-care and hospital records) with the following recorded history:
- diagnoses: none recorded
- procedures: male genital [N171] [N171] (first recorded 30-40)
- prescriptions: none recorded

Your task is to: estimate the probability that he is diagnosed with "skin disorder" for the first time in the next 3 years. He has never been diagnosed with it.

Qwen3.5-9B, no tools

0.8

Qwen3.5-9B w/ CW

<cohort-think> I need to estimate the probability of a first-time diagnosis of "skin disorder" within 3 years for a 52-year-old UK male with no prior diagnoses, but with a history of male genital procedures (N171) recorded 30-40 years ago. [...] The N171 procedure from decades ago is unlikely to be directly related to current skin disorder risk, unless there's some indirect connection I'm not aware of. [...] </cohort-think> I need to find an appropriate comparison cohort for this patient. Let me start by looking at what cohorts are available from my current position.

cw_look()

cw_look → {"state_id": "default:c6b96cf81f207d8e", "age_band": "50-54", "sex": "Male", "opcs": [N171], "n_band": "10-49", "legal_moves": {"opcs.zoom_coarser": ..., "opcs.drop": ..., "icd.add": ...}}

I can see I'm already in a cohort that matches the patient's profile exactly: [...] Now I need to observe the rate for "skin disorder" at a 3-year horizon.

cw_observe("disease": "skin disorder", "horizon": "3")

cw_observe → {"status": "RATE_WITHHELD_HERE", "disease": "L11", "horizon": 3, "hint": "This group's rates are not published. Move to a group you can read and reason from there about how this one differs.", "moves_left": 24}

The rate is withheld for this exact cohort. I need to move to a group where the rate is published and then reason about how this patient's risk differs from that group. [...] The N171 procedure is from 30-40 years ago (around ages 22-32). This is quite old and unlikely to be directly relevant to skin disorder risk in a 52-year-old. Dropping this constraint (option 2) would give me a larger cohort of males aged 50-54 with no specific procedure history, which should be a reasonable comparison group for someone with no current diagnoses.

cw_move("move_id": "opcs.drop.N171->default:c2bf5de0dceaa135")

cw_move → {"hit": true, "cost": 2.0, "state_id": "default:c2bf5de0dceaa135", "icd": [], "opcs": [], "bnf": [], "n_band": "1000+"}

Now I'm in a cohort of UK males aged 50-54 with no recorded diagnoses, procedures, or prescriptions. This is a very clean comparison group. Let me observe the rate for "skin disorder" at a 3-year horizon for this group.

cw_observe("disease": "skin disorder", "horizon": "3")

cw_observe → {"status": "published", "disease": "L11", "disease_name": "skin disorder [L11]", "horizon": 3, "rate": 0.0, "at_risk": "1000+"}

Interesting! The rate for skin disorder (L11) within 3 years for UK males aged 50-54 with no recorded history is 0.0 (zero events observed in this cohort of 1000+ people). [...] Given the data available, I believe this patient's risk is approximately equal to the cohort rate of 0.0, perhaps with a very small increase due to the unique procedural history, but likely still very close to 0%. [...]

cw_commit("rates": "[0.0, 0.0, 0.0, 0.0]")

Qwen3.5-9B w/ CW + CS-GRPO

<cohort-think>Compared with people who share the rest of this history but for whom a procedure on the male genital is not required, this patient's risk of a first diagnosis of skin disorder within 3 years is higher. The comparison group provides a reasonable baseline because their overall health profile appears similar, lacking recorded diagnoses or prescriptions that might influence skin health. The procedure on the male genital, recorded some time ago, likely represents a significant event potentially related to underlying health issues or inflammation that can sometimes manifest in the skin or increase susceptibility to skin conditions. This past intervention might also suggest a degree of altered immune function or scarring that could predispose him to developing skin problems compared to those without such a history. Therefore, we see a noticeably increased chance, perhaps roughly double the baseline risk, in this patient. Adjustment: -0.0001.</cohort-think>

You have said which group you would compare this patient against and by how much their risk differs. Now carry it out. [...]

cw_look()

cw_look → {"state_id": "default:c6b96cf81f207d8e", "opcs": [N171], ...}

cw_move("move_id": "opcs.drop.N171->default:c2bf5de0dceaa135")

cw_move → {"hit": true, "cost": 2.0, "state_id": "default:c2bf5de0dceaa135", "icd": [], "opcs": [], "bnf": [], "n_band": "1000+"}

cw_observe()

cw_observe → {"status": "published", "disease": "L11", "disease_name": "skin disorder [L11]", "horizon": 3, "rate": 0.0, "at_risk": "1000+"}

cw_commit()

Extraction. The recorded answer is $r = 0.0000$. True CM is $r^{\text{ref}} = 0.0000$ and true CC is $\text{sign}(r - r^{\text{ref}}) = 0$, indistinguishable, so the item leaves the CC denominator. No tools: CM extracted 0.0080, miss. With CohortWorld: CM extracted 0.0000, exact. After CS-GRPO: CM extracted 0.0000, exact. Both agents reach

the target cohort by the single drop of `cw_move` that removes the procedure and read 0.0000 there, the base agent after one withheld reading at the entry cohort.

A.3 COHORTS BY THE LEVEL γ

Both collections are built at three depths of the coding trees, and Figure 6 shows one real branch of each tree, ICD-10 for diagnoses, OPCS-4 for procedures, and BNF for prescriptions, at the three depths. Moving one level down the tree splits a history feature into its children, so the cohorts keyed on it get smaller and more specific. Table 5 gives all six counts. CohortWorld mixes all levels, while CohortCal, shown in Table 3, is γ=1st-depth. All levels are released.

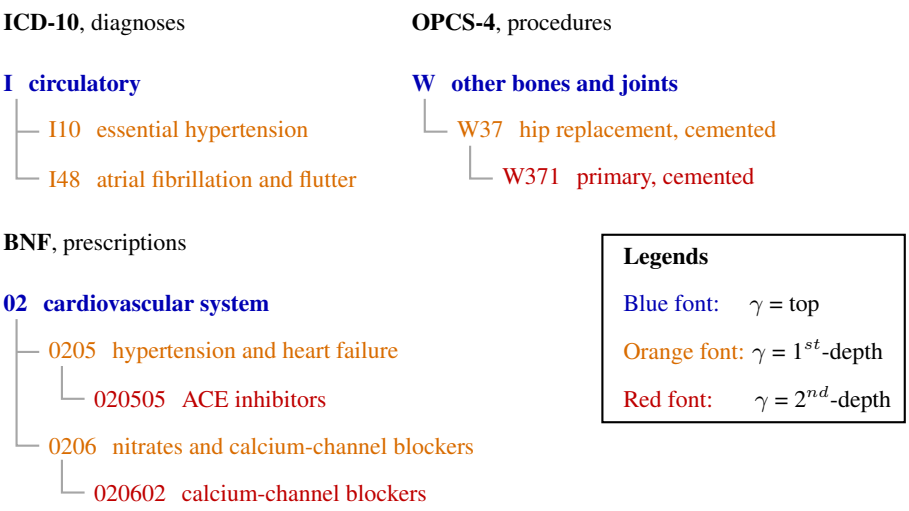

Figure 6: One example of each coding tree at the three depths of cohort key. Colour marks the depth and the lines are the parent-child structure of the code itself. A cohort at a coarser γ pools every history under the branch, so the cohort size is larger and its disease risk is steadier.

Table 5: Released cohorts at each γ level.

collection	γ=top	γ=1st-depth	γ=2nd-depth	total
CohortCal	3,640	3,198	2,866	9,704
CohortWorld	5,619	7,113	7,448	20,180
both	9,259	10,311	10,314	29,884

A.4 TRAINING DETAILS

Cold start. The first step is to finetune Qwen3.5-9B and Qwen3.6-27B on the doctor-like RWE diagnostic CoT. The demonstrations are written by code rather than by hand, so the comparison cohort named in each `<cohort-think>` block and the signed difference δ that closes it are read off the collection and are correct by construction. Table 6 gives the settings, which are the same for both sizes apart from the initial weights and the number of training GPUs. The 1,753 training samples at a batch of 48 make 36 steps an epoch, so 72 steps are exactly two epochs.

Reinforcement learning. Training runs inside CohortWorld on the training split, with the two-turn episode and the reward of Section 3.3. Table 7 gives the settings. We report two checkpoints of CS-GRPO post-training for each model size. The first takes 100 steps for Qwen3.5-9B and 40 steps for Qwen3.6-27B from the cold start with $w_\psi = 0$, so that ψ is omitted at the early stage. The second takes 50 and 20 further steps from it with $w_\psi = 0.3$.

The scale u of the ψ term is derived rather than tuned. It is the environment's own calibration scale divided by two, which gives $u = 0.0359$, so a declared difference within 0.025 of the true one earns credit above the fixed baseline of $\frac{1}{2}$, and one further away is charged. The three levels of λ pay for issuing any tool call, for reaching a cohort other than the one the agent starts on, and for reading a rate there.

Table 6: Cold start. The rendered samples have a median length of 303 tokens and a maximum of 486 over a 500-sample probe, which is what sets the sequence length.

setting	Qwen3.5-9B	Qwen3.6-27B
initial weights	Qwen3.5-9B	Qwen3.6-27B
training samples	1,753	1,753
steps	72	72
batch size	48	48
micro-batch size	1	1
sequence length	1,024	1,024
learning rate	1×10^{-5}	1×10^{-5}
seed	42	42
training GPUs	4xH100	8xH100

Table 7: CS-GRPO. M pseudo-subjects and N responses each form the matrix of Section 3.3. The batch counts episodes, not matrices. A cell with two values reads Qwen3.5-9B / Qwen3.6-27B.

setting	CS-GRPO	CS-GRPO w/ ψ
initial weights	cold start	$w_\psi = 0$ at step 100 / 40
steps	100 / 40	50 / 20
M, pseudo-subjects per matrix	16	16
N, responses per pseudo-subject	8	8
batch size in episodes	512	512
comparison sets per episode	8	8
smallest scorable population	4	4
replay buffer	128	128
w_ϕ, format	0.2	0.2
λ rungs, call, move, read	$0.1, 0.2, 0.3$	$0.1, 0.2, 0.3$
w_g, population term	1.0	1.0
weights inside g, AUROC and slope	$1.0, 1.0$	$1.0, 1.0$
slope target	1.0	1.0
w_ψ, declared difference	0	0.3
u, scale of ψ	–	0.0359
baseline of ψ	–	0.5
cross entropy on tool results	0.1	0.1
completion tokens	2,048	2,048
sequence length	16,384	16,384
episodes in flight	26	22
training GPUs	4 / 6xH100	4 / 6xH100
inference GPUs	2xH100	1 / 2xH100
checkpoint interval	25 / 10	25 / 10
seed	42	42

A.5 Inference details

This section describes how every model was tested.

Prompts. Below is the system prompt for CohortCal benchmarking. It is given to a model that answers with no CohortWorld tools.

> You are a clinical epidemiology estimator. For every question output ONLY your single best point-estimate percentage as a bare number (e.g. 2.3), using your general knowledge of population disease incidence. Always give a number even under high uncertainty. Never refuse, never explain, never add any words or symbols other than the number.

Single-turn models on CohortCal. The 11 open-weight models and the two closed models answer the 31,784 unseen items with only the above prompt as the system prompt. The reply is parsed as a percentage. The two closed models are served by their vendors with default sampling, no system prompts and low thinking effort.

Agents on CohortCal and on the three CohortWorld tasks. An agent receives the question prompt in the examples shown in Appendix A.2, the four tools implemented by prime-rl Intellect (2025), and a budget of 26 tool calls.

A.6 METRICS IN EXPERIMENTAL RESULTS

Output of an episode in CohortWorld RL environment. An agent ends an episode by calling **`cw_commit`** with a list of predicted disease risks, and the answer scored is their median. The environment reads the answer from the first well-formed source among that list, an `ANSWER:` line in the final message, the last rate the agent read, and a fixed report, and it records which of the four it used. A single-turn model has no episode, and its answer is the percentage parsed from its reply.

CohortMatch. For CohortMatch, the CM acc. is defined as

$$\mathrm{CM} = \frac{1}{|I|} \sum_{i \in I} \mathbf{1}\big[\hat{r}_i = r_i^{\mathrm{ref}}\big], \tag{5}$$

over the list of scored items I.

CohortContrast. For CohortContrast, the CC acc. is defined as

$$\mathrm{CC} = \frac{1}{|I'|} \sum_{i \in I'} \mathbf{1}\big[\,\mathrm{sign}(\hat{r}_i - r_i^{\mathrm{ref}}) = \mathrm{sign}(r_i - r_i^{\mathrm{ref}})\big], \qquad I' = \{\, i \in I : r_i \neq r_i^{\mathrm{ref}} \,\}, \tag{6}$$

with sign taking the three values $-$, 0 and $+$. Items with $r_i = r_i^{\mathrm{ref}}$ are the INDISTINGUISHABLE class of Section 3.2 and are excluded from the denominator.

CohortThink and CohortCal. Both score a predicted rate against a recorded one and use the same three numbers. Item i carries a truth $r_i \in [0, 1]$, a prediction $\hat{r}_i \in [0, 1]$, and a threshold $b_i > 0$, the incidence of that item's disease and horizon in the population named in the caption of the table. Items with $b_i = 0$ are dropped, since the score below is undefined there. Write

$$\ell_i = \mathbf{1}[\,r_i > b_i\,], \qquad s_i = \hat{r}_i / b_i, \tag{7}$$

so the label asks whether this history raises the risk above the disease's own prevalence and the score is the prediction in the same units. Let $P = \{i : \ell_i = 1\}$ and $Q = \{i : \ell_i = 0\}$. The three numbers are

$$\mathrm{AUROC} = \frac{1}{|P||Q|} \sum_{i \in P} \sum_{j \in Q} \Big(\mathbf{1}[s_i > s_j] + \tfrac{1}{2}\, \mathbf{1}[s_i = s_j] \Big), \tag{8}$$

$$\text{beats } b = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\big[\,|\hat{r}_i - r_i| < |b_i - r_i|\,\big], \qquad \hat{\beta} = \frac{\sum_i (r_i - \bar{r})(\hat{r}_i - \bar{\hat{r}})}{\sum_i (r_i - \bar{r})^2}. \tag{9}$$

AUROC asks only whether the model orders two people correctly and is unchanged by any monotone rescaling of $\hat{r}$. The percentage that beats b asks the opposite question item by item, whether the answer is closer to the truth than the population incidence is, so a model that reports b itself scores zero on it. The slope $\hat{\beta}$ is the ordinary least squares coefficient of $\hat{r}$ on r, and it is 1 for a model whose answers move with the truth one for one and near 0 for a model that answers the same number everywhere.

Intervals. Every interval in this paper is a percentile bootstrap over items with 2,000 resamples at seed 42, reported as the 2.5th and 97.5th percentiles of the resampled statistic.