TOWARD VIRTUAL PATIENTS: A BENCHMARK FOR FORECASTING FUTURE RADIOLOGY REPORTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrospective electronic health records (EHRs) contain outcomes only for the tests clinicians historically ordered, so interactive clinical agents cannot be trained or evaluated on queries that depart from recorded trajectories. We study an evaluable surrogate—*query-conditioned clinical world modeling*: given a patient's leakage-controlled EHR context and a requested imaging study, generate the study's report, scored by its structured findings under a frozen extractor. Building on MIMIC-IV and MIMIC-IV-Note, we align longitudinal clinical events with 117K radiology reports across four modalities and pose 1,500 sealed *masked-observed* queries (every target study was actually performed, so every query is gradable). Evaluating frontier LLMs, structured supervised pipelines, and a budgeted LLM-policy search over model design, we find a robust dissociation: end-to-end LLMs reach micro-F1 ≈ 0.55 while structured predictors over a prior-imaging finding-state representation plateau at ≈ 0.50, invariant to search budget, policy strength, and rendering strategy. We trace the plateau to the supervision itself: report labels measure radiologists' *mention behavior*, not patient state—a finding documented as present goes unmentioned in the next report $\sim 51\%$ of the time even within seven days, nearly independent of elapsed time. Under an *omission-tolerant* reading of the same predictions—forgiving a predicted finding that prior reports document and the target report merely leaves unmentioned—all systems gain 0.10–0.12 and the verdict changes: a structured pipeline that significantly trails the best end-to-end LLM under the mention convention ties it under that reading—the evaluation convention, not only the model, determines the conclusion. We formalize the task as state estimation composed with mention selection and intervene on both factors, obtaining a *double dissociation*: mention-trained models win under the mention convention while densely state-supervised models win under the omission-tolerant reading—where a dense-supervised world model nearly matches the best end-to-end LLM—and, at matched context budget, structured state supplements lift an end-to-end report generator by $+0.02$. Our model-design search runs under an integrity-by-construction protocol: the LLM policy acts in a typed, budgeted action space with the sealed test structurally unreachable; one committed configuration is scored once by an external evaluator. Validation-to-sealed gaps grow with search intensity while sealed scores stay flat, and policy behavior (early saturation, cost-insensitivity, avoidance of semantic actions) is reported as part of the evaluation. We release the benchmark, findings schema, extraction pipeline, environment, and evaluation harness.

1 INTRODUCTION

Interactive clinical agents—systems that decide which test to order next, or simulate a patient's course under a plan—need an environment that can answer clinical queries. Retrospective EHRs are the only environment available at scale, and they have a structural flaw: they contain outcomes only for the tests clinicians historically chose to perform. An agent trained against replayed records can exploit the existence of a recorded result as a signal of what clinicians suspected, and the record cannot answer queries that depart from the historical trajectory. Frontier EHR agents already hit this wall: MIRA, autonomously ordering tests inside a sandboxed MIMIC-IV record, must exclude cases whose guideline-required results were never recorded (Ferber et al., 2026). Progress on interactive

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.

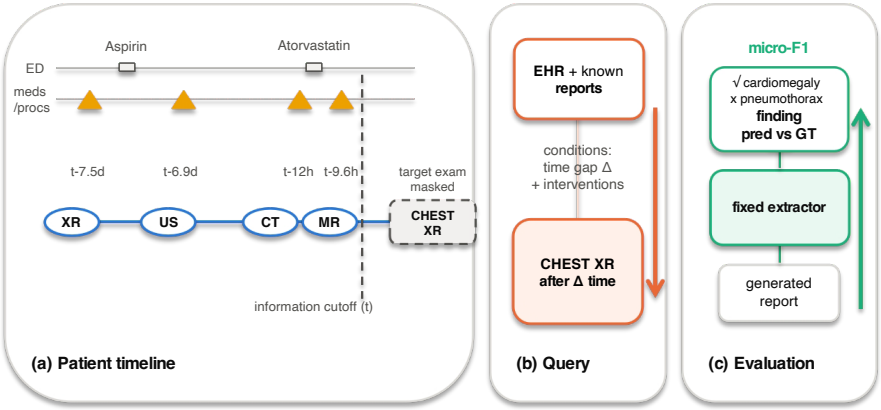


Figure 1: Query-conditioned clinical world modeling: the target study is masked; only pre-query information is visible; generated reports are scored by a frozen findings extractor against the true report.

clinical AI is therefore gated on a modeling question: can we predict what a requested examination *would show*, from the information available before it?

We take an evaluable step toward this goal. We formulate **query-conditioned clinical world modeling**: given a patient's leakage-controlled EHR context before time t and a requested imaging study, generate the study's report, scored by the structured findings it expresses under a frozen extractor. Because outcomes exist only for performed studies, we pose only such queries—a **masked-observed** surrogate that is empirically gradable and lets us measure, rather than assume away, clinician-driven selection.

Our headline result is a diagnosis, not a leaderboard. Frontier LLMs reading the full context reach micro-F1 $\approx$ 0.54–0.56; a broad family of structured predictors over a prior-imaging finding-state representation—including a budgeted LLM-policy search over model design and a conditional transition ("world model") arm—plateaus at $\approx$ 0.50, robustly across search budget, policy strength, agent memory, and rendering strategy. Tracing errors to their source shows why: **report labels measure mention behavior, not patient state**. A finding documented as present goes unmentioned in the next report $\sim$51% of the time even within seven days, nearly independently of the time gap—a silence the label cannot distinguish from resolution—and most false positives of the best end-to-end model are clinically reasonable carry-forwards the target report simply does not restate. Under an omission-tolerant reading of the same predictions—one that forgives exactly such carry-forwards—all systems gain 0.10–0.12 **and the verdict changes**: the structured pipeline's best variant significantly trails the best end-to-end LLM under the mention convention yet ties it under an omission-tolerant reading—the evaluation convention, not only the model, determines the conclusion. We package this as a task decomposition—report prediction = state estimation $\circ$ mention selection—and act on both factors: densified concurrent supervision on the training side (§4.4) produces a double dissociation across the two conventions (mention-trained wins the mention metric; state-trained wins the omission-tolerant reading, where a dense-supervised world model nearly matches the best end-to-end LLM), and a context-matched rendering control reverses an apparent negative result: at equal context budget, structured state supplements *help* ($+0.02$) (§5.5).

A second thread concerns *how* the model configurations were found. We let an LLM policy search a typed action space of clinical modeling decisions (families, ensembles, calibration, hyperparameters, state-representation subsets, transition-model knobs) under a fixed budget, with the sealed test **structurally unreachable**: no action returns sealed feedback, and one committed configuration is scored once by an external evaluator. This turns evaluation integrity into environment design and honesty into a measured quantity: validation-to-sealed gaps grow with search intensity while sealed scores stay flat, and policy behavior—saturation within $\sim$12 steps, a cheap policy matching a $50\times$ more expensive one, both avoiding high-variance semantic actions they explicitly recognize as available—is reported as part of the evaluation.

Contributions. (1) A quantified diagnosis of report-supervised clinical prediction—mention $\neq$ state—with a dual-convention metric under which model comparisons change (§5.3). (2) A benchmark instrument: 117K aligned MIMIC-IV radiology reports, a namespaced four-state findings schema with a frozen extractor, and 1,500 sealed leakage-controlled queries, with human-verification and confidence-faithfulness audits (§3). (3) An honest auto-research protocol—typed actions, budgets, evaluator-held sealed split—with an empirical account of LLM-policy search behavior (§4.3, §5.6). (4) Supervision- and rendering-side interventions targeting the two factors of the decomposition (§4.4, §5.5). We release the benchmark, schema, pipeline, environment, and harness.

2 Related Work

LLM agents for ML experimentation. MLAgentBench (Huang et al., 2024) and MLE-bench (Chan et al., 2025) evaluate agents that write and run ML code, the latter with hidden Kaggle labels graded outside the agent's container; DS-Agent (Guo et al., 2024), Data Interpreter (Hong et al., 2025), AutoKaggle (Li et al., 2024), SELA (Chi et al., 2024), and AIDE (Jiang et al., 2025) add reasoning, decomposition, and tree search over code space; classical AutoML (Feurer et al., 2015; Olson et al., 2016) is the ancestor. These systems establish LLM policies as plausible ML-design searchers but expose broad free-code action spaces and evaluate generic engineering performance. Our agent instead operates over *typed clinical model-design actions* under a budget, with the sealed clinical split unreachable from the action space throughout search—removing even the submission-feedback channel that leaderboard protocols retain. LLM-Select (Jeong et al., 2025) selects features from names alone; we place representation selection inside a sealed world-model environment and report when policies decline to use it.

Clinical world models and EHR simulation. EHRWorld (Mu et al., 2026) simulates hospitalization trajectories (notably, its zero-shot frontier baseline also plateaus near F1 0.55—on lab-abnormality labels, machine-recorded ground truth without reporting noise); Foresight (Kraljevic et al., 2024) forecasts coded concepts; MeWM (Yang et al., 2025) synthesizes post-treatment tumor imaging; synthetic-EHR generators (Theodorou et al., 2023; Wang & Sun, 2022; Biswal et al., 2021; Choi et al., 2017), fixed-endpoint benchmarks (Wornow et al., 2023; Chen et al., 2023), and sepsis RL simulators (Kiani et al., 2019; Komorowski et al., 2018) round out the space. None, to our knowledge, conditions only on pre-exam EHR context plus a requested exam and time to generate and grade the future radiology report, and clinician-driven selection is rarely the object of analysis; our masked-observed design makes it measurable.

Radiology report evaluation. CheXpert/CheXbert established condition-level labeling with uncertainty (Irvin et al., 2019; Smit et al., 2020); RadGraph and RadGraph-XL provide entity–relation schemas across modalities (Jain et al., 2021; Delbrouck et al., 2024); RadCliQ, GREEN, RaTEScore, and LLM judges align metrics with clinical error severity (Yu et al., 2023; Ostmeier et al., 2024; Zhao et al., 2024; Wang et al., 2024). We inherit this line as a *readout*: a frozen four-state namespaced extractor applied identically to gold and generated reports. Unlike image-conditioned report generation—including EHR-augmented variants that still consume the target image (Nicolson et al., 2025)—our inputs exclude the target study entirely; and where this literature treats extraction fidelity as the question, we additionally treat the *mention convention itself* as a measurable source of supervision noise.

Verbalized confidence. LLMs can express calibrated confidence in some settings (Kadavath et al., 2022; Lin et al., 2022), but reliability depends on prompt, task, and model (Geng et al., 2024; Groot & Valdenegro-Toro, 2024; Yang et al., 2024). We audit stated confidence against extractor-scored outcomes and find generation quality and confidence faithfulness dissociate.

Benchmark integrity. MLE-bench's private grading (Chan et al., 2025), RE-Bench's budgeted R&D tasks (Wijk et al., 2025), SWE-bench Verified/Live (OpenAI & SWE-bench Team, 2024), reward-hacking benchmarks (Thaman, 2026), and contamination studies (Xu et al., 2024) motivate treating agent evaluation channels as attack surfaces. We adopt integrity by construction—an action space with no sealed-evaluator channel—plus reported honesty diagnostics.

3 BENCHMARK

3.1 TASK DEFINITION

Let a patient's record be a time-ordered sequence of clinical events (labs, vitals, medications, procedures, ED events, imaging reports). At time t a clinician orders an imaging study $q = (\text{exam}, \text{modality}, \text{region}, t)$; the task is to generate, from information strictly before t, the free-text report that would plausibly be documented for q, scored by the structured findings it expresses.

Two properties distinguish this from trajectory forecasting: the task is **query-conditioned** (answer a specific requested study at a specific time, the interaction shape an interactive agent requires) and **masked-observed** (we pose only queries whose outcome was recorded, so every query is gradable). We claim nothing about studies never performed; the surrogate inherits—and lets us analyze—clinician-driven selection (§5.3).

3.2 CONSTRUCTION FROM MIMIC-IV

We derive the benchmark from MIMIC-IV and MIMIC-IV-Note (Johnson et al., 2023), aligning longitudinal structured events with 117K radiology reports (radiography, CT, MRI, ultrasound; 25 modality$\times$region namespaces). Episodes are linked across admissions with graded link confidence; each query gets a **leakage-controlled context**—prior admissions and verbatim reports, diagnoses, medications, procedures, and ED events strictly before the target time—timestamp-checked and audited (discharge summaries: prior admissions only, filtered on both chart and store timestamps), with the target study and all later information removed.

3.3 FINDINGS SCHEMA AND SCORING

A fixed extraction pipeline (gpt-5.4-mini with a frozen prompt) maps any report—gold or generated—into a namespaced findings representation: for each schema concept of the report's namespace, one of four states (*present* / *absent* / *uncertain* / *not_mentioned*). The evaluation schema is derived from the training side only and contains 73 union concepts. The primary metric treats a concept as positive if extracted as present or uncertain, and computes instance micro-F1 between generated and true positive sets, with bootstrap confidence intervals; per-namespace macro-F1 is a secondary diagnostic.

Three design decisions matter for interpretation. (i) **Every method is scored through the same frozen extractor on the same artifact type** (a free-text report), so extraction bias affects all systems identically; absolute values are extractor-relative. (ii) Gold labels are silver—extracted from the true report—audited with a 100-case stratified human verification (§3.5). (iii) The schema scores finding **presence**, not quantitative attributes (present in 27.7% of target reports; 82% of abdominal CT, 13% of chest radiographs): measurement prediction without the target image is largely ill-posed and left as an extension.

3.4 SPLITS AND SEALED PROTOCOL

The sealed test set is a proportionally namespace-stratified sample of 1,500 queries. Because the evaluation schema is train-derived only (concept vocabularies must not be fitted on test), nine low-frequency namespaces (81 queries, e.g. neck/spine CT) have no schema concepts: generated but not scorable, leaving 1,419 scored queries; both counts are always reported. The remaining 5,590 queries form the training pool; the two sets are **patient-disjoint**—zero subject and zero admission overlap—so no sealed query is a longitudinal neighbor of any training query. Queries require at least one linked prior related admission (episode continuity), which is the entire gap between the 117K aligned reports and the 7,090 queries (Appendix A). The internal validation split is drawn from the pool with the **same stratification protocol** as the test draw—the two splits independently follow one sampling rule, rather than validation being fitted to test. No predictor, hyperparameter choice, or agent-driven search may touch the sealed set: it is scored once per committed configuration by the evaluator (§4.3). This is structural, not conventional: the search agent's action space contains no operation that returns sealed-set feedback.

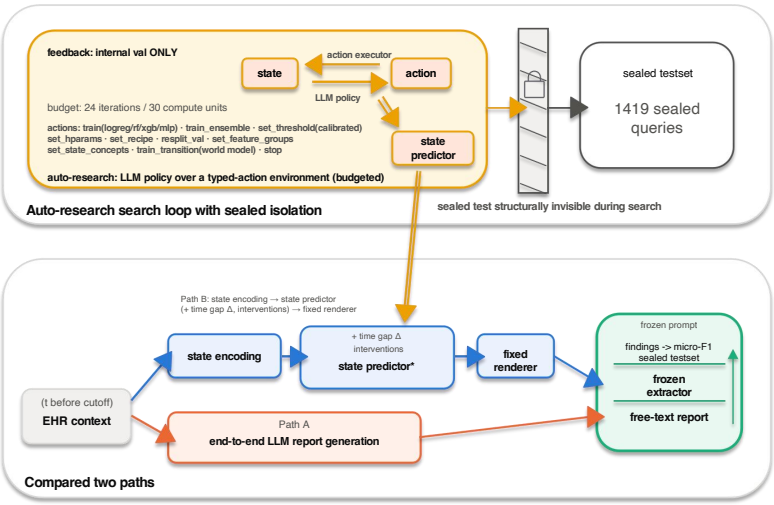


Figure 2: Both paths emit the same artifact, scored by the same frozen extractor on the same sealed queries; the auto-research policy's typed, budgeted action space has no channel to the sealed evaluator.

3.5 Trustworthiness analyses

Human verification (n=100). A stratified sample of sealed cases was audited by an author (LLM-assisted) against the true report text and coded diagnoses. Coverage: 60% full / 31% partial / 9% poor; contradictions: 85% none / 7% minor / 8% major; spurious extractions: 92% none / 8% some. Excluding the eight major-contradiction cases changes every headline micro-F1 by ≤ 0.0002. Re-extracting the audited reports with a stronger LLM (gpt-5.6-sol; 89.8% agreement) fixes none of the flagged majors: the dominant failure is schema normalization, not comprehension—the synonym-mapping layer is the indicated fix. **Independent annotation.** Two radiologists (plus a third, LLM-assisted annotator, disclosed and reported separately) independently re-audited the same 100 cases on 5-point scales: mean coverage 4.4–4.7/5 with poor coverage (≤ 2) in 1–9% of cases and explicit contradiction (≥ 4) in 0–3% (LLM-assisted: 7%); within-one-point cross-rater agreement 0.81–0.99—corroborating the author audit (Appendix H).

Self-reported confidence faithfulness. End-to-end models classify their confidence in each generated report into fixed verbal classes. Comparing stated confidence with realized per-case F1: gpt-5.4-mini's low-confidence classes are true warnings (micro-F1 0.12–0.21 versus its 0.555 average); gpt-6-astra is moderately discriminative (0.41 at "less than even"); claude-opus-4-7 is nearly unfaithful—cases it marks "chances are slight" score 0.48, indistinguishable from its average. Generation quality and confidence faithfulness dissociate (Figure 5 (right)).

4 Methods and Evaluation Framework

4.1 Two report-producing paths, one scoring channel

Path A (end-to-end). A single LLM receives the full leakage-controlled context and the query and writes the report directly (gpt-6-astra, gpt-5.4-mini, claude-opus-4-7, glm-5.3-flash).

Path B (structured-state pipeline). The context is first encoded into a structured *imaging finding state*—not a full patient state: a vector $s(t^-) \in \mathcal{S}^{73}$, $\mathcal{S} = \{\text{present, absent, uncertain, not_mentioned}\}$, holding for each union concept the value carried by the patient's prior reports (extracted per-report with the same frozen pipeline, then aggregated by deterministic recency override). Labs, vitals, diagnoses, and notes are not encoded; Path B's input is a strict subset of Path A's. A **state predictor** f_θ maps this state and mandatory conditions to

per-concept positive probabilities for the target study, $\hat{p} = f_\theta(s(t^-), \Delta, a) \in [0,1]^{|\mathcal{C}_{\text{ns}}|}$, with Δ the bucketized time gap and a the typed interventions (medications, procedures, ED events); thresholds calibrated on internal validation binarize $\hat{p}$ into the predicted state. A fixed renderer (gpt-5.4-mini, frozen prompt) writes a report expressing the predicted state, which enters the same extractor-based scoring as Path A. Both paths emit the same artifact and are scored identically (Figure 2).

4.2 State-predictor arms

Per-namespace ML ladder. Independent per-concept classifiers per target namespace (logreg, RF, XGBoost, MLP, soft-vote ensembles)—a deliberately query-conditioned baseline.

Conditional transition model (world-model arm). A single model over the full 73-concept state space: $s(t^-)$ plus conditions $\rightarrow s(t) \in \mathcal{S}^{73}$, supervision masked per case to observed concepts. Learners: a torch MLP (width/depth/dropout, concat-vs-FiLM injection) and pooled per-concept classifiers over the same input/output space. A $\Delta{=}0$ identity self-supervision option anchors continuity (mixing ratio a bounded knob).

LLM-as-state-module. An LLM performs the state update itself (state dict in/out), isolating "LLM as transition function" from Path A's free-form generation.

Per-concept decision thresholds for all arms are calibrated on internal validation only.

4.3 Budgeted auto-research with a sealed evaluator

We treat model configuration as a search problem executed by an LLM policy over a **typed action environment**: train a family or ensemble, toggle threshold calibration, set bounded hyperparameters, re-split validation, select feature groups or an explicit concept subset for the *input evidence* (with an optional output veto), train the transition arm with menu-valued knobs, or stop. Every action is typed and validated—no free code—and every training action costs budget (24 iterations, 30 compute units per run). The policy sees only internal-validation feedback; **the sealed set is structurally unreachable**: no action returns sealed scores, and the single committed configuration is scored once, outside the loop, by the evaluator—strengthening hidden-label protocols (Chan et al., 2025) by removing the submission-feedback channel entirely during search. Identical protocols run with policies of different strength; we report cost, validation-to-sealed gaps, and policy behavior—honesty of the search as a measured quantity.

4.4 Supervision-noise fix: dense concurrent state labels

Motivated by §5.3, we densify **training** supervision only (window reports supply labels, never inputs): a concept's label at time t is the state asserted by the nearest mentioning report within ± 3 days (target report has final say); the per-case mask expands to all concepts observed by any window report; and definite-prior entries that remain unmentioned leave the loss (silence is not a transition). Validation keeps the single-report mention convention—the sealed protocol—so threshold calibration, not the training target, owns the state-to-mention mapping. Dense labeling triples definite supervision per case ($3.4 \rightarrow 9.5$) and cuts the not_mentioned share from 73% to 38%.

5 Experiments

All systems are scored on the same 1,419 scorable sealed queries with the same frozen extractor; we report instance micro-F1 with 95% case-level bootstrap CIs of approximately ± 0.016 for all Table 1 rows, and paired case-level bootstrap tests for every between-arm comparison we interpret (Appendix E). Path-B thresholds are calibrated on internal validation only.

5.1 Main results

Table 1 summarizes. Two no-learning baselines anchor the table: copying the most recent same-type prior report (0.495) and carrying the prior state forward (0.474). The structured plateau clears pure persistence by at most 0.03 and end-to-end models by 0.04–0.07: the differentiating skill is mention selection, not persistence. Sealed state-direct scores of all structured arms searched under

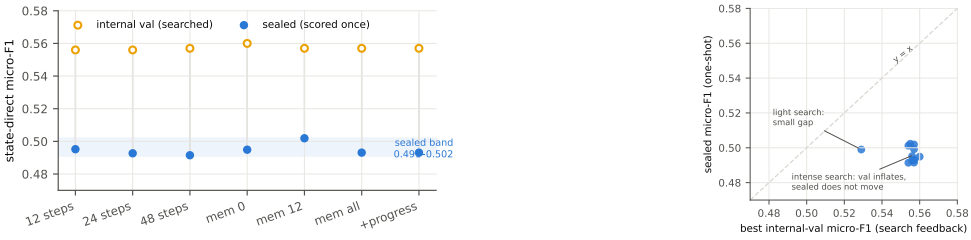


Figure 3: **Left:** across sweep cells, searched validation scores (hollow) sit 0.53–0.56 while one-shot sealed scores (filled) stay in a 0.011-wide band. **Right:** committed configurations across policies and prompts: validation inflates with search intensity; sealed does not.

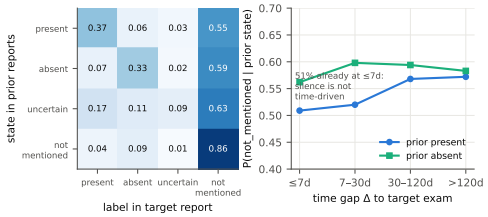


Figure 4: Report labels measure mention behavior. **Left:** row-normalized transitions from the prior-report state to the target-report label. **Right:** the present→not_mentioned rate is ∼51% already within 7 days and nearly flat in Δ: silence is not time-driven.

the mention convention (per-namespace ladder; six auto-research committed configurations; the conditional transition arm) lie within 0.491–0.506 (the best, 0.506, is a full-state pooled arm; one instructive failure is analyzed in §5.6).

5.2 The 0.50 plateau is not a search problem

Across two policies (50× cost difference), two prompt generations, and seven budget/memory sweep cells, internal-validation best configurations span 0.529–0.560 while sealed state-direct stays within 0.491–0.502. The validation-to-sealed gap *grows* with search intensity (0.030 → 0.053–0.063) while sealed scores do not move—direct evidence that (i) the sealed protocol prevented leakage and (ii) the plateau is a property of supervision and representation, not search effort.

5.3 Diagnosis: report labels measure mention behavior, not patient state

Flicker. In training pairs, a concept documented present in a prior report is labeled not_mentioned in the target report 54.5% of the time; the present→not_mentioned rate is 50.9% within 7 days and only 57.2% beyond 120 days (Figure 4). The Δ-invariant base shows silence is not time-driven, but cannot separate convention from resolution completing within days—and report-only adjudication cannot settle it: radiologists grounded their "resolved" judgments in the expectation that a persisting finding would be restated (the convention itself) and deemed persistence undecidable without images, which MIMIC-IV lacks (Appendix H). Either way, not_mentioned is not a state assertion.

Error anatomy. For the best end-to-end model, 70.5% of false positives carry definite prior-positive evidence that the target report is merely silent about; pure hallucinations are 21%. On the miss side, 65% of false negatives were never mentioned in any prior report—new findings inferable only from the indication. Path B shows the same anatomy (45–48% omission-type FPs). Three prior puzzles follow at once: naive Δ/intervention features cannot help (the label's Δ-dependence is drowned by Δ-invariant reporting noise); F1-calibrated thresholds are pushed to high recall (the model learns that known-present findings "disappear" half the time); and the plateau tracks the noise floor of mention prediction rather than a limit of state estimation.

The convention changes the verdict. We define an omission-tolerant secondary reading: a false positive is forgiven iff the concept has prior-positive evidence and the target report does not contradict it. Under this reading—same predictions, different convention—every system gains ∼0.11

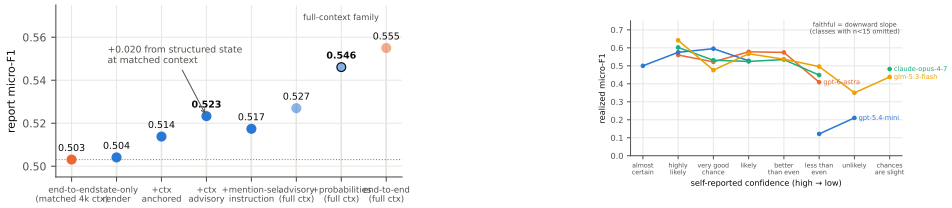


Figure 5: **Left:** the render spectrum at matched 4k context; structured supplements add $+0.020$ over the matched end-to-end control. **Right:** self-reported confidence vs realized F1; quality and faithfulness dissociate.

Table 1: Main results on the sealed set (1,419 scorable queries; micro-F1 under the frozen extractor; omission-tolerant reading in §5.3). *State-direct* rows score the predictor's binary finding-state output directly, without rendering.

	System	micro-F1	omission-tolerant
No learning	copy most recent same-type prior report	0.495	0.590
	carry forward prior finding state (state-direct)	0.474	—
Path A	gpt-6-astra (full context)	0.561	0.668
	gpt-5.4-mini (full context)	0.555	0.666
	claude-opus-4-7 (full context)	0.538	0.651
	glm-5.3-flash (full context)	0.536	0.648
	gpt-5.4-mini (4k compressed context)	0.503	0.605
LLM module	LLM-as-state-module (mini; state-direct 0.371)	0.473	—
Path B	auto-research committed → render (state only)	0.504	0.615
	+ context, advisory	0.523	0.638
	+ context, advisory, full context	0.527	—
	+ context, probabilities (mini / astra renderer)	0.546 / 0.545	0.667 / —
	full-state pooled world-model arm (state-direct)	0.506	—

(Table 1), dwarfing every between-model gap. The gaps themselves also change: Path B's best render variant (probability-annotated, full context) trails the best end-to-end model significantly under the mention convention (0.546 vs 0.561; paired bootstrap $p = 0.02$) and ties it under this reading (0.667 vs 0.668; $p = 0.95$, Appendix E); forgiveness volumes are symmetric across the two arms, so the *contrast* is meaningful even though absolute levels are inflated. The primary metric remains the mention convention (what the data can grade for every arm); the omission-tolerant reading is a heuristic sensitivity analysis that *upper-bounds* state accuracy—a bound report-based adjudication can neither validate nor refute: of 100 sampled forgiven false positives, radiologists clinically affirmed persistence for 6%, called 71% resolved on restatement-expectation grounds (the mention convention itself), and found 13% in fact mentioned by the target (extractor misses; Appendix H). The two conventions answer different questions—*what will be written* versus *what is going on*—and neither report labels nor their silence measure the latter directly.

Task decomposition. These results motivate reading the task as report prediction = finding-state estimation ∘ mention selection, with §4.4 targeting the first factor and §5.5 probing the second.

5.4 Supervision-side intervention: dense concurrent labels

Dense concurrent supervision triples definite labels and removes the silence channel (§4.4)—and *lowers* sealed scores under the mention convention (Table 4, Appendix; the per-namespace control is flat). Training toward finding state and predicting report mentions **trade off**. Two corroborations: (i) the full-state pooled arm on original labels matches the frontier (0.502): architecture is not the bottleneck; (ii) scoring the same four models under both conventions yields a **double dissociation**: with the stable pooled-XGB learner, mention-trained beats state-trained on the mention metric (0.502 vs 0.444) while state-trained beats mention-trained under the omission-tolerant reading (0.646 vs 0.630); the transition MLP shows the same crossover with larger amplitude (0.666 under that reading, approaching the best end-to-end model's 0.668). The rule: the training target must sit on the same side of the mention/state distinction as the evaluation convention—itself a choice (§5.3).

5.5 RENDERING-SIDE ANALYSIS: THE RENDER SPECTRUM

The render family shares one compressed 4,000-character context with a context-matched end-to-end control (Figure 5). **(1) The Path-A advantage is mostly context budget**: end-to-end mini drops $0.555 \rightarrow 0.503$ on it. **(2) Structured state substitutes for missing context, not full context.** At matched compressed context the advisory renderer beats the control (0.523 vs 0.503; paired $p = 0.001$). At *full* context the same supplements yield only 0.527 vs end-to-end's 0.555: the same context increase buys $+0.052$ end-to-end yet $+0.004$ through the state-scaffolded prompt. An ablation localizes that $+0.052$: removing diagnosis codes, chief complaints, triage, and discharge summaries (45% of the context) costs only -0.009 (n.s.)—the advantage is almost entirely verbatim prior reports. The interference is largely a *binarization* artifact: the predictor's probability separates re-mentioned from silent prior-positive findings with AUC 0.76—the mention-selection factor of §5.3—but F1-calibrated thresholds (0.10–0.25) discard this signal, handing the renderer flat positive lists. Passing per-finding probabilities with reading guidance instead recovers precision ($0.417 \rightarrow 0.475$) and lifts the full-context renderer to **0.546**—within end-to-end's CI. The effect is not context-dependent: a *state-only* probability render (no raw context) reaches 0.524 vs its flat counterpart 0.504, and beats the information-matched end-to-end@4k (0.503). The supplement's value also shrinks with renderer capability ($+0.020 \rightarrow -0.009 \rightarrow -0.016$ across the capability/context axis): gpt-6-astra with probabilities (0.545) stays below its own end-to-end (0.561)—a stronger reader already extracts what they encode. **(3) Prompt-level mention selection fails**: instructing the renderer to reason as the reporting radiologist does not reduce omission-type false positives ($1217 \rightarrow 1186$; overall 0.517): the factor needs signal (e.g., the indication), not instruction.

5.6 AUTO-RESEARCH BEHAVIOR

(1) *Budget and memory*: search saturates by $\sim$12 steps; memory windows $\{0, 6, 12, \text{all}\}$ are indistinguishable on sealed (0.493–0.502). (2) *Cost*: the gpt-5.4-mini policy matches gpt-6-astra at roughly $1/50$th of its per-run API cost; a uniform-random policy over the same action space commits 0.39–0.47 (Appendix F). (3) *Semantic actions*: the concept-subset action was used zero times in thirteen trajectories despite documentation and hints; the policy *recognizes* and declines it as "risky" (Figure 6). Neither policy used the free per-namespace diagnostics and knobs either; the strong policy inspected once and doubled down on the world-model arm (0.506): conservatism tracks the data's payoff structure, not feedback visibility. (4) *Policy strength changes what is explored, not what is found*: the stronger policy probed the world-model arm and once committed the configuration with the largest validation-to-sealed gap ($0.546 \rightarrow 0.450$), while the cheap policy's conservative ensembles transferred at 0.49–0.50.

6 DISCUSSION AND LIMITATIONS

What the plateau means. The plateau is not unpredictability: under the omission-tolerant reading the same pipelines score 0.61–0.67, the deficit concentrating in mention selection and context budget. **Silver ground truth; scale.** Gold findings are extracted, not hand-labeled; all systems share the extractor, so comparisons are internally consistent and extractor-relative. The 100-case audit and independent re-audit (§3.5) bound coverage, contradiction, and spuriousness. The benchmark is single-institution and chest-radiograph-heavy (63%; 73 unique concepts across 9 scorable namespaces, 1,419 sealed queries)—adequate for the analytical claims; small model differences should respect the CIs.

Finding state, not patient state. The predictor's basis (prior imaging findings plus typed conditions) is an incomplete patient description, and the under-determination is *self-flagging*—extreme probabilities (61% of pairs) are 94.9% accurate, the mid-range near chance (58.7%)—which is what makes the supplements usable; the ablation localizes the missing signal to prior-report *text*. Four low-resource namespaces (16% of queries) score zero under every configuration—a data-volume limit. Finally, the agent search found little beyond well-tuned ensembles—reported as such.

Conclusion. We turned "can a model answer clinical queries beyond the recorded trajectory?" into a gradable benchmark—which returned a sharper question (what do report labels measure?), changed verdicts, and showed model discovery can be audited honestly.

AI USE STATEMENT

(Required; does not count toward the page limit.) Generative AI systems are both the *subjects* and *instruments* of this work. As subjects, LLMs are evaluated as report generators, state-update modules, and auto-research policies; all such uses are part of the reported experiments. As instruments: (i) the findings ground truth and all scoring use a frozen LLM extractor, whose error profile we audit with a 100-case human verification (§3.5); (ii) the human audit itself was performed by an author with LLM assistance, and a stronger LLM was used to re-extract the audited subset as a robustness check; of the three independent annotators in Appendix H, one disclosed using LLM (gpt-5.6-sol) assistance during annotation—that annotator's judgments are reported separately throughout and excluded from the primary (two-radiologist) statistics; (iii) an LLM agent assisted the literature survey, and all cited claims about related work are being verified against the original papers; (iv) manuscript text and analysis code were drafted with LLM assistance and reviewed by the authors. We take responsibility for the final content, including all text, claims, and artifacts produced with AI assistance.

ETHICS STATEMENT

MIMIC-IV and MIMIC-IV-Note are de-identified and used under their data use agreements by credentialed authors. We publicly release derived schemas, code, decision traces, and aggregate results—never raw note text; all patient-level derivatives, including per-case identifiers, are confined to a PhysioNet derived-data project under the MIMIC-IV DUA. The figure example uses relative timestamps and paraphrased content. The benchmark evaluates retrospective prediction and is not a clinical decision tool; omission-tolerant readings are diagnostic conventions, not endorsements of any reporting practice.

REPRODUCIBILITY STATEMENT

The release is two-tier. The public repository carries the benchmark construction pipeline, findings schema, frozen extraction prompts, the typed auto-research environment, all decision traces, and aggregate results. All patient-level MIMIC-IV derivatives—per-case prediction records, per-case probabilities, generated reports and extractions, and sealed-set identifiers—are distributed as a PhysioNet derived-data project under the MIMIC-IV DUA, following PhysioNet policy for derived data. All LLM components are pinned by exact model identifier and frozen prompt text, so credentialed users can recompute every reported score without querying any model. Anonymous repository: `https://anonymous.4open.science/r/TVP-A-BENCHMARK-FOR-SIMULATING-DISEASE-PROGRESSION-VIA-FUTURE-RADIOLOGY-REPORT-GENERATION-6B97`.

REFERENCES

Siddharth Biswal, Soumya Ghosh, Jon Duke, Bradley Malin, Walter Stewart, Cao Xiao, and Jimeng Sun. EVA: Generating longitudinal electronic health records using conditional variational autoencoders. In Ken Jung, Serena Yeung, Mark Sendak, Michael Sjoding, and Rajesh Ranganath (eds.), *Proceedings of the 6th Machine Learning for Healthcare Conference*, volume 149 of *Proceedings of Machine Learning Research*, pp. 260–282. PMLR, 2021. URL `https://proceedings.mlr.press/v149/biswal21a.html`.

Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, Lilian Weng, and Aleksander Madry. MLE-bench: Evaluating machine learning agents on machine learning engineering. In *International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=6s5uXNWGIh`.

Emma Chen, Aman Kansal, Julie Chen, Boyang Tom Jin, Julia Rachel Reisler, David A. Kim, and Pranav Rajpurkar. Multimodal clinical benchmark for emergency care (MC-BEC): A comprehensive benchmark for evaluating foundation models in emergency medicine. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2023. URL `https://arxiv.org/abs/2311.04937`.

Yizhou Chi, Yizhang Lin, Sirui Hong, Duyi Pan, Yaying Fei, Guanghao Mei, Bangbang Liu, Tianqi Pang, Jacky Kwok, Ceyao Zhang, Bang Liu, and Chenglin Wu. SELA: Tree-search enhanced LLM agents for automated machine learning. *arXiv preprint arXiv:2410.17238*, 2024. URL https://arxiv.org/abs/2410.17238.

Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F. Stewart, and Jimeng Sun. Generating multi-label discrete patient records using generative adversarial networks. In Finale Doshi-Velez, Jim Fackler, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens (eds.), *Proceedings of the 2nd Machine Learning for Healthcare Conference*, volume 68 of *Proceedings of Machine Learning Research*, pp. 286–305. PMLR, 2017. URL https://proceedings.mlr.press/v68/choi17a.html.

Jean-Benoit Delbrouck, Pierre Chambon, Zhihong Chen, Maya Varma, Andrew Johnston, Louis Blankemeier, Dave Van Veen, Tan Bui, Steven Truong, and Curtis Langlotz. RadGraph-XL: A large-scale expert-annotated dataset for entity and relation extraction from radiology reports. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12902–12915. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.765. URL https://aclanthology.org/2024.findings-acl.765/.

Dyke Ferber, Lars Hilgers, Christiane Höper, Benedict Kinny-Köster, Jan-Niklas Eckardt, Katharina Egger-Heidrich, Marius Bill, Martin M. K. Schneider, Jan Clusmann, Lejla Kadric, Marcel Oehme, Maximilian Mayrhofer-Schmid, Alexander Oeser, Georg Wölflein, Isabella C. Wiest, Jan Moritz Middeke, A. John Iafrate, Daniel Truhn, Dirk Jäger, and Jakob Nikolas Kather. Towards autonomous medical artificial intelligence agents. *Nature*, 655(8125):1282–1291, 2026. doi: 10.1038/s41586-026-10675-5.

Matthias Feurer, Aaron Klein, Katharina Eggensperger, Jost Tobias Springenberg, Manuel Blum, and Frank Hutter. Efficient and robust automated machine learning. In *Advances in Neural Information Processing Systems*, volume 28, pp. 2962–2970, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/11d0e6287202fced83f79975ec59a3a6-Abstract.html.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6577–6595. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.366. URL https://aclanthology.org/2024.naacl-long.366/.

Tobias Groot and Matias Valdenegro-Toro. Overconfidence is key: Verbalized uncertainty evaluation in large language and vision-language models. In *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing*, pp. 145–171. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.trustnlp-1.13. URL https://aclanthology.org/2024.trustnlp-1.13/.

Siyuan Guo, Cheng Deng, Ying Wen, Hechang Chen, Yi Chang, and Jun Wang. DS-Agent: Automated data science by empowering large language models with case-based reasoning. In *Proceedings of the 41st International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2402.17453.

Sirui Hong, Yizhang Lin, Bang Liu, Bangbang Liu, Binhao Wu, Ceyao Zhang, Danyang Li, Jiaqi Chen, Jiayi Zhang, Jinlin Wang, Li Zhang, Lingyao Zhang, Min Yang, Mingchen Zhuge, Taicheng Guo, Tuo Zhou, Wei Tao, Robert Tang, Xiangtao Lu, Xiawu Zheng, Xinbing Liang, Yaying Fei, Yuheng Cheng, Yongxin Ni, Zhibin Gou, Zongze Xu, Yuyu Luo, and Chenglin Wu. Data interpreter: An LLM agent for data science. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19796–19821. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.1016. URL https://aclanthology.org/2025.findings-acl.1016/.

Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. MLAgentBench: Evaluating language agents on machine learning experimentation. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings*

of the 41st International Conference on Machine Learning, volume 235 of *Proceedings of Machine Learning Research*, pp. 20271–20309. PMLR, 2024. URL https://proceedings.mlr.press/v235/huang24y.html.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):590–597, 2019. doi: 10.1609/aaai.v33i01.3301590. URL https://doi.org/10.1609/aaai.v33i01.3301590.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curtis P. Langlotz, and Pranav Rajpurkar. RadGraph: Extracting clinical entities and relations from radiology reports. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. URL https://arxiv.org/abs/2106.14463.

Daniel P. Jeong, Zachary C. Lipton, and Pradeep Ravikumar. LLM-select: Feature selection with large language models. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=xINiaZIrob.

Zhengyao Jiang, Dominik Schmidt, Dhruv Srikanth, Dixing Xu, Ian Kaplan, Deniss Jacenko, and Yuxiang Wu. AIDE: AI-driven exploration in the space of code. *arXiv preprint arXiv:2502.13138*, 2025. URL https://arxiv.org/abs/2502.13138.

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. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023. doi: 10.1038/s41597-022-01899-x. URL https://doi.org/10.1038/s41597-022-01899-x.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022. URL https://arxiv.org/abs/2207.05221.

Amirhossein Kiani, Chris Wang, and Angela Xu. Sepsis world model: A MIMIC-based OpenAI gym "world model" simulator for sepsis treatment. *arXiv preprint arXiv:1912.07127*, 2019. URL https://arxiv.org/abs/1912.07127.

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. doi: 10.1038/s41591-018-0213-5. URL https://doi.org/10.1038/s41591-018-0213-5.

Zeljko Kraljevic, Dan Bean, Anthony Shek, Rebecca Bendayan, Harry Hemingway, Joshua Au Yeung, Alexander Deng, Alfred Balston, Jack Ross, Esther Idowu, James T. Teo, and Richard J. B. Dobson. Foresight–a generative pretrained transformer for modelling of patient timelines using electronic health records: A retrospective modelling study. *The Lancet Digital Health*, 6(4):e281–e290, 2024. doi: 10.1016/S2589-7500(24)00025-6. URL https://doi.org/10.1016/S2589-7500(24)00025-6.

Ziming Li, Qianbo Zang, David Ma, Jiawei Guo, Tuney Zheng, Minghao Liu, Xinyao Niu, Yue Wang, Jian Yang, Jiaheng Liu, Wanjun Zhong, Wangchunshu Zhou, Wenhao Huang, and Ge Zhang. AutoKaggle: A multi-agent framework for autonomous data science competitions. *arXiv preprint arXiv:2410.20424*, 2024. URL https://arxiv.org/abs/2410.20424.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=8s8K2UZGTZ.

Linjie Mu, Zhongzhen Huang, Yannian Gu, Shengqian Qin, Shaoting Zhang, and Xiaofan Zhang. EHRWorld: A patient-centric medical world model for long-horizon clinical trajectories. *arXiv preprint arXiv:2602.03569*, 2026. URL https://arxiv.org/abs/2602.03569.

Aaron Nicolson, Shengyao Zhuang, Jason Dowling, and Bevan Koopman. The impact of auxiliary patient data on automated chest X-ray report generation and how to incorporate it. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 177–203. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.9. URL https://aclanthology.org/2025.acl-long.9/.

Randal S. Olson, Nathan Bartley, Ryan J. Urbanowicz, and Jason H. Moore. Evaluation of a tree-based pipeline optimization tool for automating data science. In *Proceedings of the Genetic and Evolutionary Computation Conference 2016*, pp. 485–492. ACM, 2016. doi: 10.1145/2908812.2908918. URL https://doi.org/10.1145/2908812.2908918.

OpenAI and SWE-bench Team. SWE-bench Verified. https://openai.com/index/introducing-swe-bench-verified/, August 2024. Human-validated subset of SWE-bench.

Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson, Michael Moseley, Curtis Langlotz, Akshay S. Chaudhari, and Jean-Benoit Delbrouck. GREEN: Generative radiology report evaluation and error notation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 374–390. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.21. URL https://aclanthology.org/2024.findings-emnlp.21/.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew Lungren. CheXbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 1500–1519. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.117. URL https://aclanthology.org/2020.emnlp-main.117/.

Kunvar Thaman. Reward hacking benchmark: Measuring exploits in LLM agents with tool use. *arXiv preprint arXiv:2605.02964*, 2026. URL https://arxiv.org/abs/2605.02964. Accepted to ICML 2026.

Brandon Theodorou, Cao Xiao, and Jimeng Sun. Synthesize high-dimensional longitudinal electronic health records via hierarchical autoregressive language model. *Nature Communications*, 14(1):5305, 2023. doi: 10.1038/s41467-023-41093-0. URL https://doi.org/10.1038/s41467-023-41093-0.

Zifeng Wang and Jimeng Sun. PromptEHR: Conditional electronic healthcare records generation with prompt learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 2873–2885. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.185. URL https://aclanthology.org/2022.emnlp-main.185/.

Zilong Wang, Xufang Luo, Xinyang Jiang, Dongsheng Li, and Lili Qiu. LLM-RadJudge: Achieving radiologist-level evaluation for x-ray report generation. *arXiv preprint arXiv:2404.00998*, 2024. URL https://arxiv.org/abs/2404.00998.

Hjalmar Wijk, Tao Lin, Joel Becker, Sami Jawhar, Neev Parikh, Thomas Broadley, Lawrence Chan, Michael Chen, Josh Clymer, Jai Dhyani, Elena Ericheva, Katharyn Garcia, Brian Goodrich, Nikola Jurkovic, Holden Karnofsky, Megan Kinniment, Aron Lajko, Seraphina Nix, Lucas Sato, William Saunders, Maksym Taran, Ben West, and Elizabeth Barnes. RE-Bench: Evaluating frontier AI R&D capabilities of language model agents against human experts. In *International Conference on Machine Learning*, 2025. URL https://arxiv.org/abs/2411.15114.

Michael Wornow, Rahul Thapa, Ethan Steinberg, Jason A. Fries, and Nigam H. Shah. EHRSHOT: An EHR benchmark for few-shot evaluation of foundation models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2023. URL https://openreview.net/forum?id=iEfqvyp9w3.

Cheng Xu, Shuhao Guan, Derek Greene, and M-Tahar Kechadi. Benchmark data contamination of large language models: A survey. *arXiv preprint arXiv:2406.04244*, 2024. URL https://arxiv.org/abs/2406.04244.

Daniel Yang, Yao-Hung Hubert Tsai, and Makoto Yamada. On verbalized confidence scores for LLMs. *arXiv preprint arXiv:2412.14737*, 2024. URL https://arxiv.org/abs/2412.14737. Revised in 2026.

Yijun Yang, Zhao-Yang Wang, Qiuping Liu, Shuwen Sun, Kang Wang, Rama Chellappa, Zongwei Zhou, Alan Yuille, Lei Zhu, Yu-Dong Zhang, and Jieneng Chen. Medical world model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8319–8329, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Yang_Medical_World_Model_ICCV_2025_paper.html.

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y. Ng, Curtis P. Langlotz, Vasantha Kumar Venugopal, and Pranav Rajpurkar. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9):100802, 2023. doi: 10.1016/j.patter.2023.100802. URL https://doi.org/10.1016/j.patter.2023.100802.

Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. RaTEScore: A metric for radiology report generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15004–15019. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.836. URL https://aclanthology.org/2024.emnlp-main.836/.

A Benchmark statistics

The sealed set stratifies 1,500 queries over 18 namespaces (chest radiography 60%, other radiography 13%, head CT 8%, head MR 4%, chest/abdominal CT 8%, ultrasound 4%, other 3%); the training pool holds 5,590 queries. The gap between the 117K aligned reports and the 7,090 queries is the episode-continuity requirement: a query is posed only for a target study whose patient has at least one linked prior related admission, since disease-progression forecasting is undefined without a documented prior trajectory; the remaining reports still contribute as context, extraction-schema sources, and prior-state evidence. Numeric measurements appear in 27.7% of target reports (abdominal CT 82%, chest CT 75%, head CT 57%, ultrasound 58%, chest radiographs 13%). Intervention vocabularies: 60 medications, 60 procedures (train-derived). **Same-study leakage audit.** A preliminary read of the target study stored as an earlier note would leak the answer into the context. Scanning all 1,500 sealed queries: 6 have a same-exam-name report within one hour before the target; text comparison shows every one is a distinct acquisition (repeat portable radiographs around interventions; indications differ; max text similarity 0.60), not a preliminary read of the target study—legitimate pre-exam information. Full per-namespace tables and episode-length distributions ship with the public release; per-case audit CSVs are in the DUA-gated component.

B Per-namespace results

Sealed micro-F1 by namespace (mention convention; scorable queries only). Small-n rows carry wide intervals and should not be ranked. Three observations: (i) head CT is hard for *every* arm, including frontier LLMs (0.11–0.17)—the benchmark's headroom is not chest-radiograph saturation; (ii) the structured probability render is competitive off-chest (best arm on ultrasound and abdominal CT) while trailing on non-chest radiography, where heterogeneous exam types defeat its fixed schema; (iii) the two smallest namespaces score zero for every arm, the data-volume limit noted in §6.

Table 2: Per-namespace sealed micro-F1 (report-level, frozen extractor).

Namespace	n	astra	mini	Path-B prob.	advisory@4k	state-only	copy prior
chest radiography	900	0.588	0.584	0.571	0.553	0.529	0.540
other radiography	127	0.625	0.455	0.286	0.000	0.000	0.083
head CT	115	0.170	0.109	0.154	0.106	0.131	0.104
chest CT	60	0.603	0.638	0.630	0.576	0.575	0.357
ultrasound (other)	59	0.167	0.154	0.500	0.000	0.000	0.000
abdominal CT	57	0.458	0.420	0.474	0.415	0.357	0.359
head MR	54	0.180	0.189	0.220	0.191	0.138	0.154
abdominal radiography	28	0.000	0.000	0.000	0.000	0.000	0.000
extremity radiography	19	0.000	0.000	0.000	0.000	0.000	0.000

C EXTRACTOR SENSITIVITY

Re-extracting a stratified 400-case subset—gold reports plus generated reports from both key arms, which span the two generation styles (free end-to-end LLM text vs. structured rendering)—with a stronger extractor (gpt-5.6-sol in place of the frozen gpt-5.4-mini) changes micro-F1 by ≤ 0.007 per arm and the between-arm gap by ≤ 0.009, well within sampling error: synonym-normalization idiosyncrasies act symmetrically on gold and predictions under a shared extractor and cancel in aggregate, empirically supporting the frozen-extractor design.

D BOOTSTRAP CONFIDENCE INTERVALS

95% case-level bootstrap (2,000 resamples) for Table 1: gpt-6-astra $0.561\,[0.545, 0.578]$; gpt-5.4-mini $0.555\,[0.540, 0.571]$; claude-opus-4-7 $0.538\,[0.520, 0.554]$; gpt-5.4-mini@4k $0.503\,[0.487, 0.519]$; Path-B state-only $0.504\,[0.489, 0.519]$; Path-B advisory $0.523\,[0.508, 0.539]$. glm-5.3-flash $0.536\,[0.519, 0.553]$.

E PAIRED SIGNIFICANCE TESTS

Case-level *paired* bootstrap for the between-arm comparisons interpreted in the text: cases resampled with replacement (B=10,000) over the 1,419 scorable sealed queries, micro-F1 difference recomputed per resample; two-sided percentile p. Differences the text interprets as real are significant here; differences the text describes as ties are not. Per-case records for every arm ship with the DUA-gated component of the release, so credentialed users can recompute all rows offline. Two robustness notes on the omission-tolerant comparison: (i) forgiveness volumes are symmetric across the tied arms—59.7% of gpt-6-astra's false positives are forgiven vs 59.9% of Path-B's—so the tie is not an artifact of one arm differentially exploiting the forgiving rule; (ii) the copy-prior baseline has the *highest* forgiveness share (74.4%) yet the lowest omission-tolerant score (0.590): the reading is not a persistence subsidy.

F AUTO-RESEARCH ENVIRONMENT

Typed actions with argument schemas and budget costs: `train(method)` (1 unit), `train_ensemble([methods])` (n units), `train_transition(config)` (2 units; bounded menus for hidden width, depth, dropout, condition injection, self-supervision ratio, learner), `set_threshold_mode`, `set_recipe`, `set_hparams` (bounded ranges), `set_min_pos`, `resplit_val`, `set_feature_groups`, `set_state_concepts(names, mode)` (validated against the concept vocabulary; evidence vs. tied-state semantics), `stop`. Off-menu arguments are rejected with an error observation. The environment exposes remaining budgets, current configuration, best-so-far, and the most recent action results; sealed access is absent from the action set. Full sweep tables (budget $\times$ memory $\times$ progress; validation and sealed scores) and all decision traces are released.

Table 3: Paired bootstrap on sealed micro-F1 differences (positive = first arm higher).

Comparison	Convention	Δ	95% CI	p
Path-B advisory@4k − end-to-end mini@4k	mention	$+0.020$	$[+0.008, +0.033]$	0.001
end-to-end mini − copy prior report	mention	$+0.060$	$[+0.043, +0.078]$	<0.001
astra − Path-B probabilities (full ctx)	mention	$+0.015$	$[+0.002, +0.028]$	0.020
astra − Path-B probabilities (full ctx)	omission-tolerant	$+0.000$	$[-0.014, +0.014]$	0.95
Path-B probabilities − end-to-end mini	mention	-0.009	$[-0.020, +0.002]$	0.12
Path-B probabilities − end-to-end mini	omission-tolerant	$+0.001$	$[-0.011, +0.014]$	0.83
astra − mini	mention	$+0.006$	$[-0.006, +0.019]$	0.31
mini − mini (context ablation)	mention	$+0.009$	$[-0.002, +0.020]$	0.12
opus − mini	mention	-0.017	$[-0.030, -0.005]$	0.007
opus − mini	omission-tolerant	-0.014	$[-0.029, +0.000]$	0.054
glm − mini	omission-tolerant	-0.017	$[-0.031, -0.003]$	0.015
Path-B state-only render − copy prior report	mention	$+0.009$	$[-0.010, +0.029]$	0.35
Path-B state-only render − copy prior report	omission-tolerant	$+0.026$	$[+0.004, +0.048]$	0.021

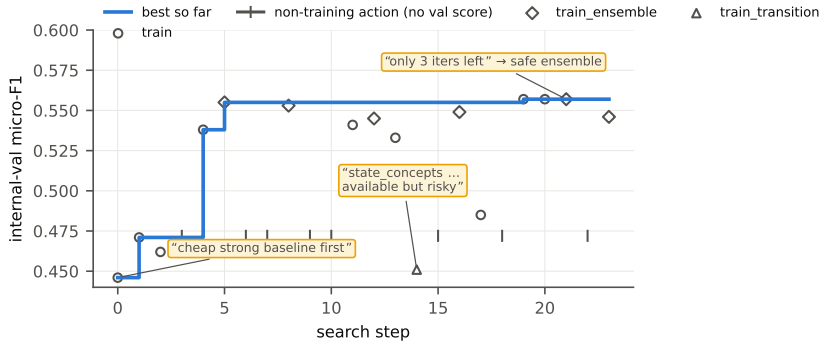


Figure 6: A representative auto-research trajectory (24 steps): family comparison, calibration, one world-model probe, then ensemble tuning; callouts quote the policy's reasoning, including risk-averse avoidance of the concept-selection action.

Torch transition-MLP runs show CPU-nondeterministic run-to-run variance (flagged wherever quoted); pooled sklearn learners are seeded.

Random-policy control. A uniform-random policy over the same action space, budgets, and commit rule (best internal-validation configuration) was run with three seeds: internal-validation bests 0.475–0.488 and sealed state-direct **0.394 / 0.426 / 0.469**—all below every LLM-policy committed configuration (0.491–0.506) and, in two of three seeds, below even the carry-forward baseline (0.474). LLM policies therefore perform genuine search—they find better configurations than chance under identical budgets—while the ≈ 0.50 ceiling is a property of the supervision, which no policy pierces. The random-policy script and traces are released.

G Human verification protocol

Stratified 100-case sample (seed fixed); the annotator (an author, LLM-assisted) sees the true report text, coded diagnoses from the pre-exam context, and the extracted four-state attributes for the target namespace, and answers: coverage of key findings (full/partial/poor), contradictions (none/minor/major, with concepts), spurious extractions (none/some/severe). Numeric-only omissions are recorded separately and do not count against coverage.

H Independent annotation protocol and results

Three annotators worked independently on identical items in identical order, in a self-contained offline interface; one annotator disclosed LLM (gpt-5.6-sol) assistance and is reported separately (see

Table 4: Supervision gate: sealed state-direct micro-F1 (mention convention) when the *training* labels move from single-report mentions toward finding state. The per-namespace arm uses unchanged binary labels (control).

Training labels	transition (MLP)	pooled XGB	pooled logreg	per-ns XGB (control)
single-report mentions (orig)	0.462	**0.502**	0.434	0.484
+ silence unsupervised (maskA)	0.376	0.445	0.390	0.484
+ dense $\pm$3d concurrent (denseB)	0.345	0.444	0.403	0.484

AI use statement). **Task A (gold-label verification; the same 100 stratified sealed cases as §3.5).** Per case: true report text, the extractor's four-state gold attributes, and 1–5 scales for coverage, contradiction, and spuriousness. Two radiologists: coverage means 4.7 and 4.4; poor coverage ($\leq$2) 1% and 9%; explicit contradiction ($\geq$4) 0% and 3%; spuriousness means 1.07 and 1.27. LLM-assisted annotator: 4.4 / 9% / 7% / 1.11. Cross-rater exact agreement 0.42–0.88, within-one-point 0.81–0.99; chance-corrected κ is depressed by the skew toward "no issue" ratings. Radiologist medians collapsed to the author audit's 3-way categories agree at 0.54 (coverage), 0.77 (contradiction), 0.90 (spuriousness). **Task B (omission adjudication).** 100 forgiven false positives sampled from the union of three arms (probability render, astra, mini), blinded among 20 attention controls (10 target-explicit-absent, 10 target-explicit-present); per item: the prior report asserting the finding, the target report, the time gap, and a 1–5 persistence scale plus two exception flags ("target actually mentions it", "prior label unsupported"). Radiologists passed 8–10/10 of each control type. On forgiven items (median of the two radiologists; flagged items excluded): 71% likely resolved ($\leq$2), 23% indeterminate, 6% likely persistent ($\geq$4); 13% majority-flagged as actually mentioned by the target (frozen-extractor mention misses) and 8% as unsupported prior labels; between-radiologist linearly weighted $\kappa = 0.32$. The resolved judgments concentrate in fast-resolving findings (atelectasis, opacities, pulmonary edema; median gap 2.9 days). The LLM-assisted annotator diverged (46% flagged as mentioned; 64% of its scored items judged persistent) and is excluded from the primary statistics; its flags typically cite implicit or synonymous mentions (e.g., *midline shift* for mass effect), an interpretive disagreement about what counts as a mention that again implicates the synonym-normalization layer (§3.5). **Debriefing.** Asked for their judgment basis and then re-confirmed per item, the radiologists reported that on the forgiven items *every* 1–2 score was grounded in the expectation that a persisting abnormality would have been restated by the reporting radiologist—i.e., the reporting convention itself—whereas 4–5 reflected clinical reasoning (e.g., chronicity); the single clinically grounded low score was an attention-control item whose target report explicitly negates the finding, precisely the case the forgiveness rule already excludes. They considered reliable persistence adjudication to require the underlying images, which MIMIC-IV does not include; a per-item judgment-basis table derived from this debriefing ships with the release. The adjudication therefore bounds what reports alone can decide rather than providing image-level state ground truth—itself a demonstration of the mention/state gap: even experts reading both reports cannot separate state from convention. Consequences for the omission-tolerant reading are discussed in §5.3; the annotation-interface builder, the analysis script, and answers keyed by opaque item ids ship with the public release; answers keyed by case identifiers, the item mapping, and the rendered interfaces (which embed MIMIC-derived report text) are in the DUA-gated component.

I CONFIDENCE-FAITHFULNESS TABLES

Per-class micro-F1 (count) underlying Figure 5 (right); glm-5.3-flash per-class values are in the released records.

Stated confidence	gpt-6-astra	gpt-5.4-mini	claude-opus-4-7
Almost certain	0.71 (9)	0.50 (15)	—
Highly likely	0.56 (274)	0.58 (479)	0.60 (179)
Very good chance	0.52 (64)	0.60 (230)	0.53 (148)
Likely	0.58 (660)	0.53 (656)	0.52 (466)
Better than even	0.57 (278)	0.00 (2)	0.53 (528)
Less than even	0.41 (133)	0.12 (17)	0.45 (42)
Unlikely	—	0.21 (18)	—
Chances are slight	0.00 (1)	0.00 (1)	0.48 (56)