

Benchmarking Large Language Models for Adaptive Information Acquisition and Diagnosis with Sequentially Revealed Inpatient Records

Huawei Feng¹ Dongsheng Shi¹ Linlin Wang^{1,2*}

¹East China Normal University ²City University of Hong Kong
{hwfeng, dongsheng}@stu.ecnu.edu.cn llwang@cs.ecnu.edu.cn

Abstract

While recent medical benchmarks for Large Language Models (LLMs) have begun to incorporate clinical temporality, they still overlook a key reality: clinical evidence is encountered incrementally, requiring repeated decisions between reviewing the next available record and committing to a prediction. To address this issue, we introduce INA-Bench¹, a benchmark for Inpatient Note Acquisition that evaluates LLMs’ ability to make acquire-or-predict decisions using sequentially revealed real-world Chinese inpatient records. In this benchmark, each case sequentially reveals up to three note blocks: an admission note, a first-course summary, and a daily-progress summary. After each block, the model under evaluation must either request the next block or predict one or more later documented follow-up actions (laboratory, imaging, pathology, or consultation). Surgery and discharge records are withheld for label construction and boundary checks only. INA-Bench is built from 8,986 de-identified inpatient records collected from two tertiary teaching hospitals, with patient-disjoint and temporal evaluation splits. Representative LLMs exhibit diverse acquire-or-predict behaviors. For example, DeepSeek-chat improves action micro-F1 by +0.102 and +0.074 on the patient-disjoint and temporal splits, respectively, while reading fewer note blocks, though the gains weaken under temporal drift.

1 Introduction

Most existing medical benchmarks evaluate Large Language Models (LLMs) only after the relevant evidence in a patient chart has been fully assembled. However, this setting differs from prospective inpatient care, where clinical records accumulate over time and decisions often must be made before later notes become available. In practice,

clinicians must determine whether the evidence currently available is sufficient to act or whether further information is needed. For example, an admission note for a patient with fever and cough may report ground-glass opacities and suspected viral pneumonia, prompting consideration of laboratory follow-up or specialist consultation, while later evidence may support or alter that decision.

Deciding when to stop gathering evidence and act is a classical clinical decision problem. Pauker and Kassirer (1980) cast it as a threshold beyond which further testing should give way to a decision, and stopping rules have since been derived for sequences of diagnostic tests (Sonnenberg, 2002). Acquiring more evidence is therefore not automatically better, since the pursuit of certainty past the point of decision is a recognized failure mode in its own right (Kassirer, 1989). We formulate this sequential decision as **acquire-or-predict**: whether to acquire additional evidence or commit to a prediction.

Existing benchmarks do not directly evaluate this acquire-or-predict capability. Chart-prediction tasks built on MIMIC-III/IV (Johnson et al., 2016, 2023) hand the model a patient record that is already complete, and question answering grounded in clinical notes assumes the same (Pampari et al., 2018), while exam-style corpora test medical knowledge with no chart at all (Jin et al., 2021; Pal et al., 2022). Benchmarks aimed at bedside decisions still reveal the full context up front (Liu et al., 2024; Ouyang et al., 2024), as do few-shot suites scored over pre-extracted patient features (Wornow et al., 2023). Retrieval-augmented medical question answering does let a model fetch evidence, but it prices literature passages rather than notes the hospital has yet to write (Xiong et al., 2024).

A second line of work lets a model seek what it lacks: ranking clarification questions before answering (Rao and Daumé III, 2018), interviewing a patient until the case is clear (Li et al., 2024b), rais-

* Corresponding author.

¹<https://github.com/hwfengcs/INA-Bench>

ing case-specific follow-up questions (Wong et al., 2026), or refining a query across retrieval rounds (Xiong et al., 2025). Their costly unit, however, is a question, a snippet, or a simulated dialogue turn. None of them evaluates acquire-or-predict decisions over sequentially revealed inpatient records.

To address this gap, we introduce INA-Bench, an Inpatient Note Acquisition benchmark that evaluates how LLMs decide between acquiring additional evidence and committing to a prediction as real-world Chinese inpatient records are revealed sequentially. Built from 8,986 de-identified inpatient records from two tertiary teaching hospitals, INA-Bench reveals up to three note blocks per case in a fixed order: an admission note, a first-course summary, and a daily-progress summary. After each block, the evaluated model either requests the next block or commits to a multi-label prediction of follow-up actions documented later in the chart. Surgery and discharge records are never shown to the model and are used only for label construction and boundary checks. The task therefore targets subsequently documented clinical actions rather than retrospective diagnosis. Unlike conventional clinical benchmarks, which primarily score final predictions from fixed inputs, INA-Bench jointly evaluates prediction quality and acquisition behavior, including the number of note blocks read and the point at which the model commits, while also reporting the number of backend calls. In this way, INA-Bench captures not only what a system predicts, but also when it stops acquiring evidence and commits to a prediction. Representative LLMs exhibit diverse acquire-or-predict behaviors on our benchmark. Using DeepSeek-chat as the backend, the acquire-or-predict policy improves action micro-F1 over a full-visible static reader by +0.102 on the patient-disjoint split and +0.074 on the temporal split, while reading fewer note blocks but requiring more LLM inference calls.

In summary, our contributions are threefold:

- **A real-world benchmark for inpatient acquire-or-predict decisions.** INA-Bench is built from 8,986 de-identified Chinese-language inpatient records collected from two tertiary teaching hospitals. Each case reveals up to three note blocks in a fixed order. Surgery and discharge records are withheld for label construction and boundary checks. Both patient-disjoint and temporal evaluation splits are provided.
- **Controlled framework for separating acquisition**

from prediction. Our framework compares passive readers at each reveal depth, non-adaptive acquisition controls, and acquire-or-predict policies. All systems are evaluated with the same underlying LLM, reveal space, output schema, and thresholding procedure. This controlled design isolates acquisition and stopping behavior while jointly measuring prediction quality, note-block consumption, stopping depth, and LLM inference cost. Anti-copy and independent-target audits further examine robustness under alternative target definitions.

- **Extensive evaluations of decision behavior across LLMs.** We have conducted extensive evaluations across diverse LLMs, revealing substantial differences in acquire-or-predict behavior. Selective acquisition can outperform a full-visible static reader while consuming fewer note blocks, but the advantage weakens under temporal drift and may reverse when models commit prematurely. These behavioral differences cannot be captured by endpoint-only evaluation.

2 The Proposed INA-Bench

2.1 Benchmark Overview

Figure 1 provides an overview of INA-Bench. Each case consists of one inpatient record revealed block by block. Up to three *visible* note blocks are presented in a fixed order: (1) the admission note, (2) the first-course summary, and (3) a daily-progress summary. After each revealed block, the policy must either *request* the next available block or *commit* to a multi-label prediction over four documented follow-up action categories: laboratory review, imaging review, pathology review, and consultation. After the final block, the policy must commit. Surgery and discharge records are never shown to the policy. INA-Bench therefore evaluates when a policy should continue acquiring evidence and when it should commit to a prediction, rather than scoring a one-shot prediction from fully assembled evidence.

2.2 Corpus Construction

Sources and de-identification. INA-Bench is built from multi-source de-identified Chinese-language inpatient exports drawn from two tertiary teaching hospitals. De-identification is applied at the source, before the corpus reaches benchmark construction: patient names, direct identifiers, and protected health information are removed, and the re-

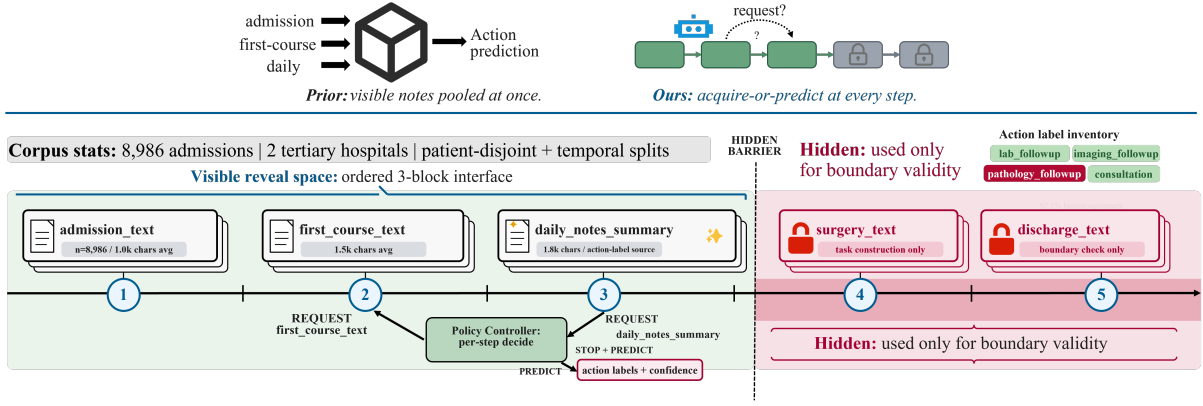


Figure 1: INA-Bench at a glance. **Top:** prior clinical benchmarks score a model over evidence that has already been pooled, whereas INA-Bench reveals one inpatient trajectory block by block and prices each reveal. **Bottom:** the reveal rail. Three blocks are visible in a fixed order and, after any prefix, the policy either requests the next block or commits a multi-label action set; surgery and discharge records are never requestable and serve only for label construction and boundary checks.

Statistic	Count
<i>Corpus and official splits</i>	
Aligned inpatient trajectories	8,986
Patient-disjoint train / dev / test	6,202 / 920 / 1,864
Temporal train / dev / test	6,290 / 898 / 1,798
Cases with daily progress notes	3,907
Cases with surgery records	4,282
Cases with postop. first-course notes	3,392
<i>Documented-action positives</i>	
Lab follow-up	1,917
Imaging follow-up	1,823
Pathology follow-up	817
Consultation	394
<i>Case mix (share of 8,986 admissions)</i>	
Emergency admission route	16.5%
Recorded surgery	47.7%
Neoplasms	16.1%
Circulatory	12.9%
Digestive	11.9%
Genitourinary	11.7%
Symptoms and signs	11.0%
<i>Coarse surgery-group mix</i>	
None	4,704
Other	1,458
Urology	1,133
General GI	641
Neuro / orthopedic	375
Gynecology	336
Hepatobiliary	243
Special procedure	96

Table 1: Statistics, official split sizes, and case mix for INA-Bench. The five percentages under *case mix* are the largest admission-ICD chapters. The *none* row counts admissions without a surgery label; the rows below it summarize the surgery-positive mix.

maining institution-specific prefixes survive only as opaque source codes that we use as a coarse provenance proxy and never as a hospital identity.

Alignment and inclusion. Records arrive as separate document streams, so construction begins by aligning admission, first-course, daily-progress, surgery, and discharge documents onto a single admission identifier. Alignment decides what a case is: an admission enters the corpus only if it carries a parseable admission note and a resolvable admission date, since the first fixes the initial observation every policy starts from and the second makes the temporal partition well defined. Documents that cannot be attached to a surviving admission are discarded. This leaves 8,986 trajectories, of which 3,907 carry daily progress notes and 4,282 carry surgery records; the two subsets overlap, and neither is required, so the corpus keeps admissions on which the action task is uninformative rather than selecting for cases that favor acquisition.

Contents of the three blocks. The visible blocks are not three views of the same text but three stages of the record, which is what makes their order informative. The admission note states the presenting complaint, history, examination findings, and initial working diagnosis; it fixes the clinical question but rarely commits to a plan. The first-course summary is the attending’s initial assessment, and it typically names a management direction while leaving specific orders open. The daily-progress summary condenses what was actually done and observed on the ward afterwards, and it is the first block in which follow-up actions appear as recorded events rather than intentions. A policy that stops early is therefore not merely reading less text; it is deciding before the record has resolved intention into action.

Dataset composition. The corpus covers a mixed

inpatient population, not a single ward. Emergency admissions account for 16.5% of cases, 47.7% have a recorded surgery, and the largest admission chapters of the International Classification of Diseases (ICD) are neoplasms (16.1%), circulatory (12.9%), digestive (11.9%), genitourinary (11.7%), and symptoms and signs (11.0%). Documented-action positives are dominated by laboratory and imaging review, while pathology review and consultation are far rarer, so per-label support is uneven by construction. Table 1 reports the resulting statistics.

Reveal order and partitions. A single linear reveal order is a deliberate trade-off: it makes the reveal budget exact and stop/continue decisions comparable, at the cost of under-approximating the free-form navigation a real clinician would perform. We evaluate on two partitions built from the same corpus, patient-disjoint assignment by stable hashing over patient identity, which keeps a patient from appearing on both sides of a split, and temporal assignment by chronological order, which reproduces the prospective setting in which later notes do not yet exist. Manifests, prompts, and cached runs are frozen for every reported number, so any two systems are compared on identical cases.

2.3 Label Construction and Validation

Extraction. Action labels record what the doctors later wrote in the chart, not a retrospective judgement of what they should have done. The extractor reads the downstream daily progress notes, normalizes whitespace, and concatenates the free-text progress-note field, the same-day order field, and auxiliary text into one string per admission. It then matches four frozen Chinese trigger lexicons, one per action family, and marks a family positive if any of its triggers fires. Because a positive label can exist only where a downstream note is present, admissions without daily notes are all-negative by construction, which makes the daily-note-positive slice the informative regime for the action task.

Failure modes and a structural coupling. Favoring coverage over semantic normalization leaves three known failure modes: broad laboratory terms and generic “recheck” phrasing over-trigger lab follow-up; historical or planned-but-unordered mentions are counted as documented action; and heterogeneous pathology phrasing is missed even where co-occurring imaging language is captured. A further, structural issue cannot be fixed by a bet-

ter lexicon: the daily-progress summary is both the label source and the third reveal option, so the target is coupled to one of the actions being scored.

Blinded audits. We treat validation as part of construction rather than a later robustness pass; Table 2 gives both audit protocols and their outcomes. In the label-validity audit, two reviewers independently re-annotated the daily-side and discharge-side action labels on a stratified 120-case sample and a third adjudicated disagreements; raw agreement was 92.1% ($\kappa \approx 0.78$), and the adjudicated pass identified three likely false positives and three likely false negatives, implying label noise in the 4–6% range. High reviewer agreement does not imply high lexicon alignment: daily-side imaging reaches 96.7% reviewer agreement but only 47.7% alignment with the lexicon, a systematic lexicon–human gap rather than reviewer disagreement. In the acquisition-plausibility audit, two reviewers judged on 60 stratified cases whether another note should be requested after admission-only and after admission+first-course exposure; they agreed that the admission note alone was insufficient in 90.0% of cases, and agreement falls to 48.3% at the second step, where the first-course summary often already suffices.

Mitigations and release. Against it we pair the original labels with the anti-copy controls and independent-target views of §3.3, and treat no single view as deconfounded ground truth. The evaluation and label-extraction code, the frozen prompts and thresholds, and the aggregate result files behind every reported number are public, while the note-bearing records remain under institutional data governance.

2.4 Evaluation Interface and Reference Probes

Shared interface. All probes share a backend, reveal space, output schema, thresholds, and maximum step budget, so under the request/predict interface the reading policy alone determines the visible evidence and the backend-call count. Writing the visible note sequence of a case as $x = (x_1, x_2, x_3)$, a policy observes a prefix $x_{\leq t}$ and then either requests the next block or predicts surgery path, documented follow-up actions, surgery group, confidence, and rationale. Inputs stay in the original Chinese note text; all probes share Chinese clinical-workflow instructions, English schema keys, and a calibration instruction that discourages unsupported positive labels. Invalid JSON is repaired by

Research-reviewer check	n	Protocol	Main agreement result	Takeaway
Label boundary	120	Stratified across both main splits (60 patient-disjoint, 60 temporal), with per-split quotas of 15 rare-positive, 15 common-positive, 15 daily-empty, and 15 no-daily-empty cases. Two independent research-team reviewers with prior medical NLP annotation experience audited daily-side and discharge-side action labels; a third reviewer adjudicated disagreements. Reviewers were blinded to benchmark gold labels and model outputs.	Overall raw agreement 92.1%; estimated Cohen’s $\kappa \approx 0.78$ (substantial agreement by conventional annotation thresholds). Consultation was most stable; daily-side imaging showed the largest lexicon-reviewer mismatch (47.7% alignment).	Documents consistent use of the label inventory. Remaining disagreement is concentrated in daily-side imaging and a small number of discharge-side imaging/pathology cases.
Acquisition plausibility	60	Stratified across both main splits (30 patient-disjoint, 30 temporal), with per-split quotas of 8 rare-positive, 12 common-positive, 5 daily-empty, and 5 no-daily-empty cases. Two independent research-team reviewers with prior medical NLP annotation experience judged whether another note should be requested after admission-only exposure and after admission+first-course exposure; disagreements were adjudicated. Reviewers were blinded to benchmark gold labels and model outputs.	Step-1 request-more agreement 90.0%; step-2 agreement 48.3%. At step 1, the preferred next document was almost always the first-course note.	Supports the sequential reveal interface as plausible for inpatient note review, but not uniformly necessary after the first-course note. Continued daily-note requests are most justified for surgical and more complex cases.

Table 2: Blinded research-reviewer checks of label boundaries and reveal-step plausibility. The audits test the benchmark directly and complement the larger automatic validity suite.

a shared parser or mapped to a conservative default. Four per-label thresholds are tuned on the 200-case development manifest for each split and system, then frozen for the 300-case test manifest.

Reference suite. The suite spans the reading policies the interface admits. *Passive readers* consume a preselected prefix in one backend call: admission-only, admission+first-course, and the full-visible static reader that sees all three blocks. *Non-adaptive acquisition controls* spend reveals without choosing them, namely a force-first control that always takes one extra block and a random-reveal control. *Acquisition policies* make one call per stop/request decision: ReAct (Yao et al., 2023) and INA-Agent, which chooses among predicting now, requesting the first-course summary, and requesting the daily-progress summary, with the next call forced to predict after two requests. INA-Agent-R adds staged prompts and a one-line carried reflection, and we report it as an *exploratory headroom* reference rather than as our central claim.

3 Experiments and Results

3.1 Setup

Manifests and protocol. Both partitions use deterministic 70/10/20 assignment, by patient hash for patient-disjoint evaluation and by admission date for temporal evaluation; all 8,986 cases have parseable admission dates. LLM comparisons use a fixed 200-case development manifest for threshold tuning and a fixed 300-case test manifest for paired evaluation on identical cases, rather than averaging over the full splits. The official manifests are controlled stratified estimates, patient-disjoint

by surgery availability and temporal by daily-note availability, so full-split temporal drift is reported separately. Multi-manifest reruns reuse the same builder under alternate seeds: three patient-disjoint and four temporal, giving seven completed reruns across the two splits.

Backend and baselines. All main-text LLM rows use the same cached DeepSeek-V3 backend (DeepSeek-chat), output schema, decoding settings, and prediction fields, decoding at temperature 0.1 with a 700-token limit, so the frontier varies the reading policy rather than model identity. Trained logistic-regression (LR), naive-Bayes, and BERT rows measure supervised headroom, not zero-shot competitors; the LR 3-doc row is a source-aware oracle because it reads the daily-progress summary directly. Because acquisition probes spend one call per request/predict decision, active-static contrasts are budgeted policy comparisons rather than equal-call inference comparisons.

Metrics and statistics. We report action micro-F1 (μ -F1) alongside the visible-note and backend-call budgets. Paired bootstrap intervals resample aligned cases. To limit multiple testing we prespecify three primary contrasts per split, namely INA-Agent against the full-visible static reader, against the force-first control, and against the random-reveal control; significance markers report uncorrected paired 95% bootstrap confidence intervals (CIs) for these contrasts only (Koehn, 2004), and all other markers are descriptive.

3.2 Main Results

Acquiring selectively beats reading everything. Under the cached DeepSeek-chat backend, with a

Method	Patient-disjoint split			Temporal split		
	Action μ -F1	Avg. docs	Avg. calls	Action μ -F1	Avg. docs	Avg. calls
<i>Same-backend frontier and controls</i>						
Admission only	0.313	1.00	1.00	0.307	1.00	1.00
Admission + first-course	0.322	2.00	1.00	0.315	2.00	1.00
Static full visible	0.327	3.00	1.00	0.346	3.00	1.00
Active force first	0.325	2.00	1.00	0.305	2.00	1.00
ReAct	0.328	2.43	2.43	0.316	2.47	2.47
<i>Same-backend active agents</i>						
INA-Agent	0.429	2.40	2.40	0.420	2.41	2.41
INA-Agent-R (exploratory headroom)	0.562	2.43	2.43	0.452	2.49	2.49
<i>Trained references and learned/source-aware boundaries</i>						
LR admission+first [†]	0.427	2.00	0.00	0.422	2.00	0.00
LR selective [†]	0.602	2.40	0.00	0.469	2.36	0.00
LR 3-doc oracle [‡]	0.679	3.00	0.00	0.659	3.00	0.00
<i>Key displayed-score contrasts</i>						
Δ INA-Agent vs. Static full visible	+0.102*	0.60 fewer	1.40 more	+0.074*	0.59 fewer	1.41 more
Δ INA-Agent-R vs. INA-Agent	+0.133*	0.03 more	0.03 more	+0.032	0.08 more	0.08 more

Table 3: Action-centered paired-manifest estimates under a shared cached DeepSeek-chat backend. INA-Agent improves over the full-visible static reader by +0.102/+0.074 action μ -F1 while reading fewer visible note blocks but making more backend calls. * marks paired bootstrap CIs excluding zero; [†] rows are trained references; [‡] directly reads the daily-progress summary and is a source-aware boundary. Shaded cells mark the best value in each block.

fixed reveal space and output schema, INA-Agent improves action micro-F1 over the full-visible static reader from 0.327 to 0.429 on patient-disjoint evaluation (paired bootstrap $\Delta = +0.102$, 95% CI [+0.063, +0.145]) and from 0.346 to 0.420 on temporal evaluation ($\Delta = +0.074$, [+0.038, +0.112]; both CIs exclude zero). It reads fewer visible note blocks (2.40 and 2.41 against 3.00) while making more backend calls (2.40 and 2.41 against 1.00). ReAct spends a comparable number of calls without the same gain, so call count alone does not explain the contrast. The gain is task-specific: surgery-path F1 falls relative to the static reader by 0.114 and 0.062, so the finding concerns documented-action acquisition rather than uniformly better chart prediction.

The gain concentrates where the task is informative, and shrinks under temporal shift. On the daily-note-positive slice, Table 4 shows INA-Agent ahead of the passive and force-first rows while operating closer to a two-note policy than to the fully static three-note reader. The patient-disjoint active-minus-static contrast on this subgroup is +0.050 with a 95% paired CI [+0.001, +0.103] that excludes zero, but the matched temporal contrast shrinks to +0.034 with a CI [−0.012, +0.079] that crosses zero. The paired contrasts against

force-first (+0.103) and random-reveal (+0.075) remain CI-positive on the same temporal subgroup, so we read the temporal active-minus-static comparison on this subset as a stress regime rather than a headline win. All seven completed same-backend reruns yield positive deltas, averaging +0.101 patient-disjoint and +0.063 temporal, so the official-manifest result is not a single favorable draw. Figure 2 places these three views side by side: the headline gain, the same gap under stricter targets, and the per-backend stopping behavior that §4 takes up.

Trained references bound the headroom. On the original daily-derived labels, LR admission+first is essentially tied with INA-Agent, LR selective is stronger at an active-matched document budget, and LR 3-doc acts as a source-aware oracle because it reads the primary label source directly. These rows quantify learnable headroom; they do not replace the within-interface active-static comparison.

3.3 Validity and Boundary Checks

A single number over a coupled target would not be interpretable, so we re-score the same comparison under views that break different parts of the coupling. No-daily and masked-trigger controls probe direct-copy explanations; hidden-document and

Method	Patient-disjoint split			Temporal split		
	Overall Action μ -F1	daily-note-positive Action μ -F1	Avg. visible note blocks	Overall Action μ -F1	daily-note-positive Action μ -F1	Avg. visible note blocks
NB admission+first	0.381	0.478	2.00	0.364	0.429	2.00
Static full	0.327	0.539	3.00	0.346	0.535	3.00
Active force first	0.325	0.521	2.00	0.305	0.466	2.00
INA-Agent	0.429	0.590	2.40	0.420	0.569	2.41

Table 4: Action-centric analysis under the zero-shot DeepSeek-chat contract. The daily-note-positive column is the informative action subset; there, INA-Agent performs best while using a note budget closer to a two-note policy than to the fully static three-note reader. A naive-Bayes (NB) row is included as a lightweight calibrated reference; stronger supervised headroom rows appear in Table 3. Blue bold marks the best value in a column, green bold the second best; for the note-budget column, lower is better.

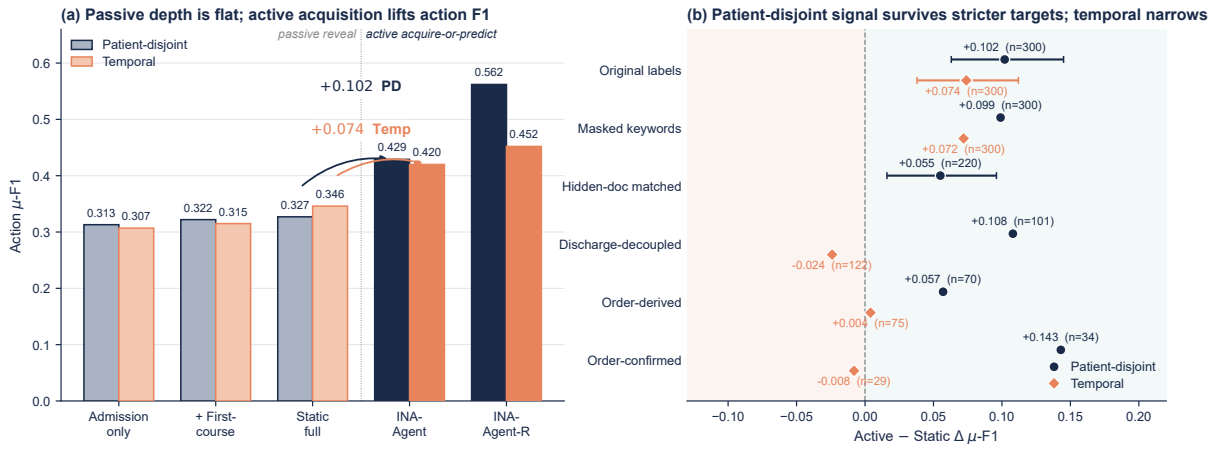


Figure 2: The headline result panel by panel, with paired bootstrap confidence intervals on the validity rows (cached DeepSeek-chat backend). **(a)** same-backend action μ -F1 against the full-visible static reader. **(b)** the active—static gap re-scored under stricter target views: positive on the patient-disjoint views, and negative at -0.024 on the discharge-decoupled temporal view. **(c)** per-backend first-step confidence against daily-note request rate, where the backends that request least are not the ones that are least confident.

discharge- or order-derived views are independent-target boundary checks; and the temporal and backend analyses of §4 delimit transfer. No single view is deconfounded, so we report the direction each one gives rather than selecting the most favorable.

Independent targets. The hidden-document annotation round covers 810 cases, and model-side comparison is limited to the 220-case matched-prediction overlap, where INA-Agent remains above the static reader by $+0.055$ μ -F1 (0.307 against 0.252; 95% CI $[+0.016, +0.096]$) and above LR selective by $+0.089$, while its gap to LR shrinks to a statistical tie. Table 5 gives the complementary discharge and order-derived view: same-backend discharge-derived targets stay positive on patient-disjoint cases ($+0.108$ on $n=101$) but turn non-positive on temporal (-0.024 on $n=122$), and order-derived subsets are directionally supportive on patient-disjoint cases and near zero under tem-

poral shift.

Anti-copy controls. Masked-trigger reruns replace the lexicon trigger strings inside the daily summary and re-tune thresholds on the masked development manifests; the paired active-minus-static delta stays positive at $+0.099$ and $+0.072$. Exact no-daily reruns hard-disable requests to the daily block, and INA-Agent still reaches 0.487/0.405 on the informative subset. Together these argue against a purely visible-keyword explanation, while also showing that part of the signal depends on later-note access. Neither removes the source coupling of §2.3.

4 Analysis: Where the Result Holds

Label-source coupling bounds the claim. The action task is most informative on the daily-note-positive slice, because positive labels are extracted from downstream daily-progress notes and cases

Table 5: Discharge-side boundary checks. The order-confirmed rows use original-label scores on cases whose action labels are corroborated by orders; they provide a corroborated subset rather than a separate target definition. Paired bootstrap CIs use 10,000 resamples.

Split	View	n	Static full	INA-Agent	INA-Agent-R	Paired Δ (INA-Agent – static)
PD	Original GT	300	0.327	0.429	0.562	+0.102 [+0.063, +0.145]*
	Discharge-derived GT	101	0.418	0.526	0.374	+0.108 [+0.053, +0.163]*
	Order-derived GT	70	0.543	0.600	0.430	+0.057 [−0.013, +0.131]
	Order-confirmed subset	34	0.571	0.713	0.714	+0.143 [+0.054, +0.234]*
Temp	Original GT	300	0.346	0.420	0.452	+0.074 [+0.038, +0.112]*
	Discharge-derived GT	122	0.532	0.508	0.404	−0.024 [−0.064, +0.014]
	Order-derived GT	75	0.670	0.674	0.515	+0.004 [−0.051, +0.060]
	Order-confirmed subset	29	0.703	0.696	0.659	−0.008 [−0.075, +0.061]

without such notes are all-negative by construction (§2.3). We therefore read INA-Agent through the validity ladder rather than as a deconfounded causal gain from acquisition. Patient-disjoint comparisons stay positive under masked triggers, no-daily controls, adjudicated hidden-document targets, and discharge-derived targets. The temporal split, by contrast, is a genuine stress regime rather than a minor robustness caveat: the share of admissions with downstream daily notes falls from 46.9% in temporal train to 30.0% in test, daily-note summaries shorten, and the discharge-derived temporal comparison becomes non-positive at −0.024. The evidence regime itself changes, so an attenuated gain there is a scope condition of the benchmark.

Backend transfer is a stopping-calibration problem. DeepSeek-chat leaves enough uncertainty for the active policy to request later notes, whereas several backends with stronger passive scores stop early and over-trust the admission evidence, which flips the active-over-static ordering outside the original setting. GPT-5-mini² requests daily notes on only 7.0% of official-manifest cases and stops after one note on 58.3%; GPT-5.4-mini and Qwen3.5-plus show the same failure at smaller but still substantial stop@1 rates. The mechanism is overconfidence, not weaker answering: GPT-5-mini stops after the admission note on 175 patient-disjoint test cases, yet exact-set accuracy on that committed slice is only 0.131 against a mean first-step confidence of 0.591, and Qwen3.5-plus has the largest confidence-accuracy gap at +0.713. INA-Bench thus separates answer quality from evidence-seeking behavior: a model can be a strong passive reader yet a weak acquisition policy, and endpoint-

only scoring would report the two identically.

Gains are uneven across clinical pathways. Stratifying the cached predictions by recorded-surgery pathway, admission route, ICD chapter, surgical specialty, and de-identified source-system proxy shows a non-uniform active-over-static gap across the case mix of §2.2. The clearest boundary is workflow-related. Under temporal shift the gap is +0.110 [+0.064, +0.158] without recorded surgery but only +0.011 [−0.048, +0.075] with it, attenuating to a tie precisely where perioperative documentation is densest. These contrasts are descriptive rather than multiplicity-corrected, the emergency cells are small ($n=54/42$), and recorded surgery, route, and source prefix are metadata proxies rather than ward, validated acuity, or hospital identity. Gains are clearest on the laboratory and imaging labels, while consultation and pathology warrant caution given their small support: pathology cases stay brittle whenever the relevant actions are phrased heterogeneously or fall outside the current coarse lexicon.

Three diagnostics must accompany the headline number. Three diagnostics follow. Under high daily-note coverage and patient-disjoint cohorts, the gap does reflect acquisition behavior. Under temporal drift or discharge-derived targets, it should be reported alongside the validity envelope rather than as a single score. Before swapping the backend, stop@1 and confidence-on-stop are worth inspecting, since an early-stopping backend turns the active-static contrast into a calibration story rather than an acquisition story.

5 Related Work

Clinical LLM evaluation under assembled evidence. INA-Bench is most directly related to clinical LLM and EHR benchmarks that score

²Backend identifiers such as GPT-5-mini, GPT-5.4-mini, and Qwen3.5-plus are vendor aliases pinned in our fixed manifests.

reasoning once the evidence has already been assembled. MIMIC-III/IV supply the de-identified records such work is built on (Johnson et al., 2016, 2023). Over those records, emrQA poses question-answer pairs against clinical notes (Pampari et al., 2018), whereas exam-style corpora test medical knowledge with no chart at all (Jin et al., 2021; Pal et al., 2022). Closer to bedside decisions, full-context clinical benchmarks report weak decision-making even when nothing is withheld (Liu et al., 2024; Ouyang et al., 2024), retrieval-augmented medical QA prices retrieved passages instead of notes (Xiong et al., 2024), and few-shot foundation-model suites score over pre-assembled features (Wornow et al., 2023). In all of them the evidence is given rather than chosen.

Asking before answering. A second line lets the system seek what it lacks: ranking clarification questions and deciding whether to ask at all (Rao and Daumé III, 2018; Shi et al., 2022; Rahmani et al., 2023), resolving ambiguity in open-domain QA (Lee et al., 2023), and planning questions by expected uncertainty reduction (Hu et al., 2024). In the clinical setting, benchmarks require a model to interview or to raise case-specific follow-up questions before committing (Li et al., 2024b; Wong et al., 2026; Gatto et al., 2025), and iterative medical RAG refines a query across retrieval rounds (Xiong et al., 2025). INA-Bench differs in unit and target: its evidence units are fixed inpatient note blocks revealed in order, and its score depends jointly on the request/predict decision, the visible-document budget, the backend-call budget, and the hidden chart documents reserved for target-side checks.

Decision-Theoretic feature acquisition. Pricing observations is an old idea. Value-of-information analysis comes from statistical decision theory (Raiffa and Schlaifer, 1961), and active learning surveys the query strategies built on it (Settles, 2009). Per-instance acquisition has since been posed jointly with classification (Shim et al., 2018; Li and Oliva, 2021) and as cost-sensitive control over which feature to buy next (Janisch et al., 2019; Kärkkäinen et al., 2019; An et al., 2022). In those settings the costly units are typically abstract features. Here they are clinical document blocks, cost is split between visible notes and backend calls, and the labels may be coupled to one of the acquirable sources, a coupling we audit explicitly.

Clinical simulation. Medical agents and virtual EHRs target broader interactive clinical outcomes, through multi-agent clinical simulation (Fan et al., 2025; Li et al., 2024a) and through agent benchmarks set inside a virtual EHR or a multimodal clinic (Jiang et al., 2025; Schmidgall et al., 2026). The concurrent ClinEnv benchmark (Lu et al., 2026) evaluates source-selective gathering and longitudinal management over MIMIC-derived admissions, querying role-specific agents across stages before submitting medications, procedures, and diagnoses. INA-Bench instead drops source selection and longitudinal management from the action space to isolate fixed-prefix stopping calibration, which is what enables paired accounting of note cost, call cost, stop position, and prediction quality. DCB (Shi et al., 2026) makes a similar move at the dialogue level, pricing conversational turns with a simulated patient rather than access to original chart documents. The supplementary material gives the detailed comparison.

6 Conclusion

INA-Bench enables the systematic evaluation of LLMs’ ability to make acquire-or-predict decisions over sequentially revealed real-world Chinese inpatient records. By jointly measuring prediction quality, evidence consumption, and stopping behavior, it captures differences that conventional endpoint-only evaluation cannot reveal. Our experiments show that sequential acquisition can improve prediction performance while requiring fewer note blocks, but the gains vary across models and diminish under temporal drift and stricter target controls. These findings highlight the importance of evaluating not only what clinical LLMs predict, but also when they choose to acquire more evidence or commit to a prediction.

Future work could extend INA-Bench to finer-grained evidence units, such as individual laboratory orders or imaging reports, to better reflect the costs of evidence acquisition in clinical practice.

Limitations

INA-Bench currently covers Chinese-language inpatient records from two tertiary teaching hospitals and uses a compact three-block reveal structure. Its results may not directly generalize to other institutions, languages, specialties, or document types. Extending the benchmark to finer-grained evidence, such as laboratory results, imaging re-

ports, or consultation notes, would require adapting the reveal space and target definitions. We use four broad documented-action categories to maintain sufficient label support and reliable evaluation across cases; finer-grained action types remain an important extension.

The fixed three-block reveal order is a controlled abstraction designed to isolate acquire-or-predict decisions rather than reproduce full clinical chart navigation. Real clinical information seeking may involve non-linear access to heterogeneous evidence sources. Extending INA-Bench to separately requestable evidence types could provide a more realistic acquisition setting and may affect acquisition behavior and relative model performance.

Our reference experiments use fixed evaluation manifests and matched underlying LLMs to enable paired and reproducible comparisons between acquisition policies. The observed benefits of sequential acquisition are not universal: they weaken under temporal drift and can disappear when models commit prematurely. Broader evaluation across additional institutions, time periods, and model families remains important future work.

Ethical Considerations

This retrospective study uses de-identified inpatient records from two tertiary teaching hospitals. All records were de-identified before benchmark construction, with direct identifiers and other personally identifying information removed. The study was reviewed and approved by the ethics committees of the participating hospitals before any experiments were conducted, and the data were used solely for research purposes. For API-based experiments, only de-identified text was transmitted, and no identifiable patient information was provided to external model providers. External API use followed the applicable institutional data-governance requirements and provider terms for research use.

Generative AI Assistance

We used generative AI tools only to improve wording and presentation. They were not used to generate hypotheses or claims, create citations, conduct analyses, or produce reported results. All content, citations, notation, tables, and numerical results were reviewed and verified by the authors against the original research records.

Acknowledgments

This work was supported by the Computational Biology Program (Grant No. 25JS2830402) of Science and Technology Commission of Shanghai Municipality (STCSM), and the Shanghai Municipal Science and Technology Major Project (Grant No. 2025SHZDZX025G06).

References

- Chaojie An, Qifeng Zhou, and Shen Yang. 2022. [A reinforcement learning guided adaptive cost-sensitive feature acquisition method](#). *Applied Soft Computing*, 117:108437.
- Zhihao Fan, Lai Wei, Jialong Tang, Wei Chen, Siyuan Wang, Zhongyu Wei, and Fei Huang. 2025. [AI hospital: Benchmarking large language models in a multi-agent medical interaction simulator](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10183–10213, Abu Dhabi, UAE. Association for Computational Linguistics.
- Joseph Gatto, Parker Seegmiller, Timothy E. Burdick, Inas S. Khayal, Sarah DeLozier, and Sarah M. Preum. 2025. [Follow-up question generation for enhanced patient-provider conversations](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25222–25240, Vienna, Austria. Association for Computational Linguistics.
- Zhiyuan Hu, Chumin Liu, Xidong Feng, Yilun Zhao, See-Kiong Ng, Anh Tuan Luu, Junxian He, Pang Wei Koh, and Bryan Hooi. 2024. [Uncertainty of thoughts: Uncertainty-aware planning enhances information seeking in LLMs](#). In *Advances in Neural Information Processing Systems*, volume 37.
- Jaromír Janisch, Tomáš Pevný, and Viliam Lisý. 2019. [Classification with costly features using deep reinforcement learning](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3959–3966.
- Yixing Jiang, Kameron C. Black, Gloria Geng, Danny Park, James Zou, Andrew Y. Ng, and Jonathan H. Chen. 2025. [MedAgentBench: A virtual EHR environment to benchmark medical LLM agents](#). *NEJM AI*, 2(9).
- Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. [What disease does this patient have? A large-scale open domain question answering dataset from medical exams](#). *Applied Sciences*, 11(14):6421.
- Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Liwei H. Lehman, Leo A. Celi, and Roger G. Mark. 2023. [MIMIC-IV, a freely accessible electronic health record dataset](#). *Scientific Data*, 10(1):1.

- Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Liwei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. [MIMIC-III, a freely accessible critical care database](#). *Scientific Data*, 3:160035.
- Kimmo Kärkkäinen, Mohammad Kachuee, Orpaz Goldstein, and Majid Sarrafzadeh. 2019. [Cost-sensitive feature-value acquisition using feature relevance](#). *arXiv preprint arXiv:1912.08281*.
- Jerome P. Kassirer. 1989. [Our stubborn quest for diagnostic certainty](#). *New England Journal of Medicine*, 320(22):1489–1491.
- Philipp Koehn. 2004. [Statistical significance tests for machine translation evaluation](#). In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.
- Dongryeol Lee, Segwang Kim, Minwoo Lee, Hwanhee Lee, Joonsuk Park, Sang-Woo Lee, and Kyomin Jung. 2023. [Asking clarification questions to handle ambiguity in open-domain QA](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11526–11544, Singapore. Association for Computational Linguistics.
- Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, and Yang Liu. 2024a. [Agent hospital: A simulacrum of hospital with evolvable medical agents](#). *arXiv preprint arXiv:2405.02957*.
- Shuyue Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S. Ilgen, Emma Pierson, Pang Wei Koh, and Yulia Tsvetkov. 2024b. [MediQ: Question-asking LLMs and a benchmark for reliable interactive clinical reasoning](#). In *Advances in Neural Information Processing Systems*, volume 37.
- Yang Li and Junier B. Oliva. 2021. [Active feature acquisition with generative surrogate models](#). In *Proceedings of the 38th International Conference on Machine Learning*, pages 6450–6459.
- Fenglin Liu, Zheng Li, Hongjian Zhou, Qingyu Yin, Jingfeng Yang, Xianfeng Tang, Chen Luo, Ming Zeng, Haoming Jiang, Yifan Gao, Priyanka Nigam, Sreyashi Nag, Bing Yin, Yining Hua, Xuan Zhou, Omid Rohanian, Anshul Thakur, Lei Clifton, and David A. Clifton. 2024. [Large language models are poor clinical decision-makers: A comprehensive benchmark](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13696–13710, Miami, Florida, USA. Association for Computational Linguistics.
- Yuxing Lu, Yushuhong Lin, Wenqi Shi, J. Ben Tamo, Xukai Zhao, Jinzhao Wang, and May Dongmei Wang. 2026. [ClinEnv: An interactive multi-stage long horizon EHR environment for agents](#). *arXiv preprint arXiv:2606.02568*.
- Zetian Ouyang, Yishuai Qiu, Linlin Wang, Gerard De Melo, Ya Zhang, Yanfeng Wang, and Liang He. 2024. [CliMedBench: A large-scale Chinese benchmark for evaluating medical large language models in clinical scenarios](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8428–8438, Miami, Florida, USA. Association for Computational Linguistics.
- Ankit Pal, Logesh Kumar Umapathi, and Malaikanan Sankarasubbu. 2022. [MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering](#). In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260.
- Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. [emrQA: A large corpus for question answering on electronic medical records](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2357–2368, Brussels, Belgium. Association for Computational Linguistics.
- Stephen G. Pauker and Jerome P. Kassirer. 1980. [The threshold approach to clinical decision making](#). *New England Journal of Medicine*, 302(20):1109–1117.
- Hossein A. Rahmani, Xi Wang, Yue Feng, Qiang Zhang, Emine Yilmaz, and Aldo Lipani. 2023. [A survey on asking clarification questions datasets in conversational systems](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2698–2716, Toronto, Canada. Association for Computational Linguistics.
- Howard Raiffa and Robert Schlaifer. 1961. *Applied Statistical Decision Theory*. Division of Research, Graduate School of Business Administration, Harvard University.
- Sudha Rao and Hal Daumé III. 2018. [Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2737–2746, Melbourne, Australia. Association for Computational Linguistics.
- Samuel Schmidgall, Rojin Ziaei, Carl Harris, Ji Woong Kim, Eduardo Pontes Reis, Jeffrey Jopling, and Michael Moor. 2026. [AgentClinic: a multimodal benchmark for tool-using clinical AI agents](#). *npj Digital Medicine*, 9(1):499.
- Burr Settles. 2009. [Active learning literature survey](#). *Computer Sciences Technical Report 1648*, University of Wisconsin–Madison.
- Dongsheng Shi, Xin Yi, Yue Li, and Linlin Wang. 2026. [Benchmarking large language models for end-to-end clinical support in traditional Chinese medicine](#). *Expert Systems with Applications*, 325:132267.

- Zhengxiang Shi, Yue Feng, and Aldo Lipani. 2022. [Learning to execute actions or ask clarification questions](#). In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2060–2070, Seattle, United States. Association for Computational Linguistics.
- Hajin Shim, Sung Ju Hwang, and Eunho Yang. 2018. [Joint active feature acquisition and classification with variable-size set encoding](#). In *Advances in Neural Information Processing Systems*, volume 31.
- Amnon Sonnenberg. 2002. [The end of diagnosis: when to stop testing](#). *The American Journal of Gastroenterology*, 97(10):2494–2498.
- Hui Min Wong, Philip Heesen, Pascal Janetzky, Martin Bendszus, and Stefan Feuerriegel. 2026. [MedClarify: An information-seeking AI agent for medical diagnosis with case-specific follow-up questions](#). *arXiv preprint arXiv:2602.17308*.
- Michael Wornow, Rahul Thapa, Ethan Steinberg, Jason A. Fries, and Nigam H. Shah. 2023. [EHRSHOT: An EHR benchmark for few-shot evaluation of foundation models](#). In *Advances in Neural Information Processing Systems*, volume 36.
- Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. [Benchmarking retrieval-augmented generation for medicine](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6233–6251, Bangkok, Thailand. Association for Computational Linguistics.
- Guangzhi Xiong, Qiao Jin, Xiao Wang, Minjia Zhang, Zhiyong Lu, and Aidong Zhang. 2025. [Improving retrieval-augmented generation in medicine with iterative follow-up questions](#). In *Proceedings of the Pacific Symposium on Biocomputing 2025 (PSB)*, pages 199–214. World Scientific.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. [ReAct: Synergizing reasoning and acting in language models](#). In *Proceedings of the Eleventh International Conference on Learning Representations*.