

THE RESEARCH LEDGER • VOLUME III

AI Agents in Healthcare Knowledge Systems

RAG, Provenance, Safety,
and Human Oversight

A Systems Engineering Treatment of Governed
Retrieval-Augmented Clinical Agents

Ngo Trung Kien, PhD

First Edition • 2026

Contents

Preface	vii
Graphical Abstract	xiii
Notation and Conventions	xv
Abbreviations	xix
List of Figures	xxvii
I Foundations	1
1 Clinical Knowledge Work and the Case for Agents	3
1.1 The Information Problem in Clinical Work	4
1.2 Four Decades of Decision Support, and What Each Generation Actually Solved	6
1.3 Medical Foundation Models: Capability Without Ac- countability	13
1.4 Retrieval as the Pivot	16
1.5 From Retrieval to Agency: Tools, Plans, and Conse- quential Action	18
1.6 Why Healthcare Breaks the Standard Agent Playbook	21
1.7 What This Book Demands of a Clinical Agent	24
2 A Taxonomy of Clinical Knowledge Agents and a Reference Architecture	33
2.1 What Kind of Object Is a Clinical Knowledge Agent?	34
2.2 Autonomy Is a Property of Actions	37
2.3 The Assurance Envelope	45
2.4 The Six-Plane Reference Architecture	50

2.5	Four Source Classes, Four Retrieval Contracts	53
2.6	Terminology Binding and Identifier-Level Grounding	57
2.7	Instantiating the Architecture	59
2.8	What the Taxonomy Rules Out	64
 II Provenance and Grounding		71
3	Clinical Knowledge Provenance and Cryptographic Lineage	73
3.1	Why Provenance Is Load-Bearing	74
3.2	A Formal Model of Clinical Lineage	78
3.3	Replay	85
3.4	Integrity: Append-Only Logs, Merkle Commitments, and Their Limits	90
3.5	Provenance and Privacy in Tension	98
3.6	The Cost of Verifiability	100
3.7	The Product Surface	102
4	Grounding, Calibration, and Principled Abstention in Clin- ical Retrieval	109
4.1	The Retrieval Plane as a Clinical Artefact	110
4.2	Retrieval Mechanics, Honestly Assessed	114
4.3	Grounding, Formalised	120
4.4	Hallucination in the Clinical Setting	129
4.5	Calibration, and Why It Does Not Travel	133
4.6	Conformal Risk Control	135
4.7	Abstention as a Decision Problem	138
 III Safety, Security, and Regulation		153
5	Threat Modelling for Clinical Knowledge Agents	155
5.1	Why the Standard Application Threat Model Is Insuf- ficient	156
5.2	Assets, Adversaries, and Objectives	161
5.3	STRIDE, Instantiated	164
5.4	Prompt Injection, Direct and Indirect	167
5.5	Data Poisoning and Corpus Integrity	175

5.6	Exfiltration and Privacy Attacks	178
5.7	Tool Abuse and Lateral Movement	180
5.8	Availability, Timing, and the Capacity to Oversee . . .	183
5.9	Failures That Behave Like Attacks	186
6	Complete Mediation, Fail-Closed Design, and Defence in Depth	193
6.1	The Security-Principles Inheritance	194
6.2	The Mediation Theorem	198
6.3	The Five Gates	206
6.4	Fail-Closed Design and Graceful Degradation	217
6.5	Defence in Depth, Quantified and Then Qualified . .	224
6.6	Output-Side Controls	227
6.7	What Guardrails Cost	229
7	Regulatory Mapping: HIPAA, FDA, ISO 14971, and the EU AI Act	237
7.1	Why Engineers Must Speak This Language	238
7.2	Privacy and Security Law as It Actually Bites	239
7.3	The Device Question	245
7.4	Risk Management as an Engineering Process	254
7.5	The EU AI Act and Its Interplay with the Device Regulation	261
7.6	The Crosswalk	264
7.7	Reporting Standards as the Bridge to Evidence	266
7.8	Distributed Responsibility and Equity	269
IV	Human Oversight and Collaboration	279
8	Human Oversight as a Budgeted Resource	281
8.1	What Oversight Is For	282
8.2	The Capacity Argument, Formalised	286
8.3	Reliance, Calibrated and Miscalibrated	296
8.4	Designing the Review Event	300
8.5	Triage: Which Actions Escalate	304
8.6	Accountability and the Moral Crumple Zone	309
8.7	Measuring Oversight	311

9	Multi-Agent Deliberation, Disagreement, and Its Limits	319
9.1	Why Anyone Would Decompose	320
9.2	Patterns and Their Clinical Instantiations	324
9.3	The Condorcet Argument and Why It Fails Here . . .	329
9.4	Disagreement Is the Product	341
9.5	New Failure Modes	347
9.6	When to Use a Panel: A Decision Rule	352
10	MediFlow: An End-to-End Governed Clinical Agent	359
10.1	The Setting: Northgate Health	360
10.2	System Composition: Six Planes, Instantiated	363
10.3	One Interaction, Eighteen Stages	370
10.4	The Latency and Cost Budget	383
10.5	Three Design Mistakes and Their Fixes	385
10.6	Evaluation Before Deployment	389
10.7	What MEDIFLOW Deliberately Cannot Do	393
V	Deployment and Practice	399
11	Algorithmovigilance: Monitoring, Drift, and Incident Response	401
11.1	Deployment Is the Beginning	402
11.2	Algorithmovigilance Defined	404
11.3	What to Instrument	405
11.4	Drift, Statistically	413
11.5	Silent Failure	424
11.6	Incident Response for Clinical AI	427
11.7	Change Management	436
11.8	Sustainability and the End of the Lifecycle	438
12	Capstone: Designing, Attacking, and Defending a Clinical Agent	445
12.1	The Brief	446
12.2	Environment, Data, and the Ethics Posture	449
12.3	Milestone 1: The Governed Baseline (Weeks 1–4) . . .	452
12.4	Milestone 2: The Red Team (Weeks 5–8)	457

12.5 Milestone 3: Remediation and the Assurance Dossier (Weeks 9–12)	463
12.6 The Rubric	468
12.7 Failure Patterns to Expect	471
12.8 Extension Projects	473
12.9 What Would Change Our Mind	474
VI Appendices	485
A The MediFlow Reference Specification	487
A.1 Component Inventory	487
A.2 Provenance Record Schema	489
A.3 Policy Rule Language	492
A.4 Deployment Manifest	494
A.5 Fail-Closed State Table	495
B Evaluation Harness and Metric Definitions	497
B.1 Retrieval Metrics	497
B.2 Grounding and Faithfulness Metrics	498
B.3 Abstention and Calibration Metrics	501
B.4 Oversight Metrics	501
B.5 Security Metrics	501
B.6 Harness Structure	502
C Red-Team Playbook for Clinical Knowledge Agents	503
C.1 Exercise Catalogue	504
C.2 Reporting Template	507
D Assurance Dossier Template	511
D.1 Dossier Contents	511
E Glossary	515
F Notes on Selected Review Questions	531
References	537
Index	575

1

Clinical Knowledge Work and the Case for Agents

Medicine has not suffered from a shortage of knowledge for at least a generation. It suffers from the cost of assembling the right fragment, about the right patient, at the moment a decision must be made. This chapter traces four decades of attempts to lower that cost, shows that each generation of decision support attacked a narrower bottleneck than the last, and argues that retrieval-augmented agents are the first technology that attacks the assembly cost directly. It then argues the uncomfortable corollary: because these systems read privileged data and can act on the world, the burden of assurance rises rather than falls.

Learning Objectives

After working through this chapter you should be able to:

- ▶ Describe the information-assembly bottleneck in clinical work and explain why it is an architectural problem rather than a content problem.
- ▶ Distinguish the four generations of clinical decision support by what each encodes, how each bounds its own behaviour, what evidence supported it, and how each characteristically fails.
- ▶ State honestly what the trial and systematic-review evidence shows about decision support changing clinician behaviour versus changing patient outcomes.
- ▶ Explain why examination-style benchmark performance in medical language models does not transfer to sequential, tool-using, patient-interactive settings.
- ▶ Name the three properties retrieval buys and the two it does not.

- ▶ Articulate the four ways in which healthcare breaks the standard autonomous-agent playbook, and connect each to one of the six planes of the reference architecture used throughout this book, distinguishing the four planes whose absence is disqualifying from the two judged on how well they are built.

1.1 The Information Problem in Clinical Work

A hospitalist admitting a patient at two in the morning is rarely stopped by ignorance of medicine. She is stopped by arithmetic. To decide whether to continue a direct oral anticoagulant she needs the current weight, the two most recent creatinine values and the interval between them, the actual dispensed dose rather than the dose someone intended, an accurate list of interacting agents including over-the-counter products, whether a planned procedure is on the schedule for the morning, and the local formulary position on dose reduction. Each of those facts exists somewhere in the record. Each lives in a different resource, under a different data model, with a different latency and a different degree of trustworthiness. The clinical question is answerable in under a minute by anyone holding the assembled facts; the assembly is what costs the minutes she does not have. The vignette carries no number, and the next paragraph says why it should not. Sixteen years ago a research agenda for the field named this problem precisely, listing the summarisation of patient-level information, the prioritisation and filtering of recommendations, and the use of free text to drive decision support among its grand challenges [195]. Those items have aged remarkably well, which is another way of saying that they were not solved.

It is worth being careful about what is and is not quantified here. The literature on documentation burden does not converge on a single canonical number: estimates of time spent in the electronic record vary by specialty, country, measurement instrument, and whether after-hours work is counted, and audit-log studies and time-motion studies routinely disagree. What is consistent is the direction and the mechanism. Clinicians spend a substantial and growing fraction of the working day in the record rather than with

patients; the fraction is larger for outpatient specialties with heavy inbox traffic; and the time is not spent reading a coherent narrative but reconstructing one. The reconstruction cost, not the reading cost, is the binding constraint. Any technology that merely presents more information more attractively addresses the wrong term in the equation.

The sedimentary record. The reason reconstruction is expensive is that the modern chart is a sedimentary deposit rather than a document. Layers accumulate and are almost never removed. Copy-forward propagates yesterday's assessment into today's note, so that a plan devised for a problem that resolved four days ago persists verbatim, gaining apparent authority with each repetition. Problem lists retain entries that were ruled out, because resolving a problem requires an affirmative act that no workflow rewards. Templates and macros insert boilerplate that satisfies billing and regulatory logic while diluting the signal-to-noise ratio of the prose around it: a normal review of systems auto-populated across fourteen organ systems is not evidence that fourteen organ systems were reviewed. Discharge summaries quote admission notes that quote referral letters. The result is a corpus in which the same assertion appears many times with different timestamps and unequal warrant, and in which recency in the document order is a poor proxy for recency in the world. **Note bloat** is usually discussed as a readability problem. For a retrieval system it is worse than that: it is a duplication attack on ranking, because the most redundant claim is the one most likely to be retrieved.

Fragmentation compounds sedimentation. The facts needed for the anticoagulation decision above are distributed across an order-entry system, a laboratory information system, a pharmacy dispensing record, a scheduling system, an external health information exchange, and the free text of notes authored by people with different training and different incentives. Each store answers a different question well and the others badly. Reconciling them requires knowing not only where a value lives but how much to trust it: an outside creatinine drawn at a dialysis unit, a weight recorded as an estimate in triage, a medication list last reconciled at a visit eight months ago. This is why the sociotechnical literature on health

information technology is so insistent that the interesting failures are failures of process — errors in entering and retrieving information, and errors in communication and coordination — rather than failures of individual data items [10]. An architecture that treats the record as a uniform document store inherits every one of those failures and adds a new one, since it will present reconciled-looking output over unreconciled inputs.

There is a final property of the clinical corpus that shapes everything in this book: it decays, and it decays silently. Guidelines are superseded, formularies change quarterly, terminology bindings drift as coding systems are revised, and locally authored rules stop firing when an upstream field is renamed. The best-documented instance of this is in decision support itself, where a case series of deployed knowledge-based content found malfunctions arising from code changes, system upgrades, and changes to external terminologies — rules that had worked for years and stopped, usually without anyone noticing until an audit [231]. Any system that answers clinical questions from a knowledge base is subject to the same rot, with the added difficulty that a language model will paper over a stale retrieval with fluent prose. Currency is therefore not a nice-to-have property of a clinical knowledge system. It is a safety property, and Chapter 4 treats it as a first-class ranking concern rather than a data-engineering afterthought.

1.2 Four Decades of Decision Support, and What Each Generation Actually Solved

The history sketched in Fig. 1.1 is often told as a story of steadily improving accuracy. That telling is misleading. Each generation of clinical decision support solved a different and narrower problem than its predecessor claimed to, and each purchased its capability with a specific, identifiable loss. Reading the history as a sequence of trade-offs rather than a sequence of improvements is the fastest way to understand why agentic systems require an assurance apparatus that none of the earlier generations needed.

Two conventions of that figure should be settled before it is used as evidence. Its upper row carries five era boxes rather than four generations, because the arrival of medical foundation models

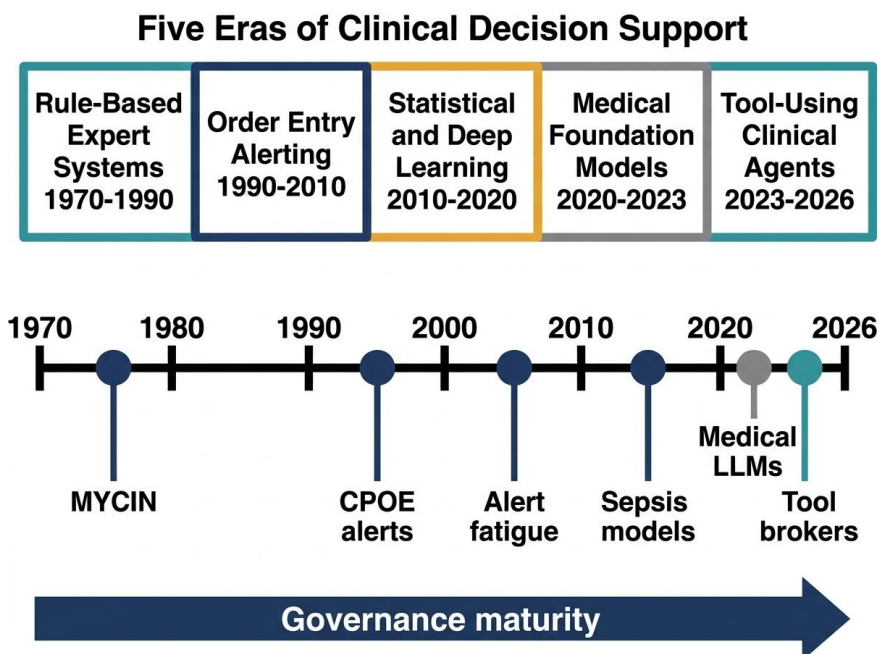


Figure 1.1. Four generations, five bands, and a governance gap. The six milestones on the lower track are identical dots on one axis — same size, same height, tinted only to pick out the two most recent eras — so the drawing says nothing about which of them reached routine clinical use — and that silence is the first thing to read from it. MYCIN, judged expert-level under blinding, and CPOE alerting, whose logic is a dose check against a table, are drawn alike, and only the second changed practice: capability and deployment are decoupled. The second thing to read is the arrow at the base, which carries no scale and is still pointing outward at the right-hand edge. Governance maturity is drawn there as a direction of travel rather than an achieved level, and it becomes binding in the last band, where a system both reads privileged data and acts on it. The drawing is illustrative throughout: the band boundaries, the dot placements and the arrow encode the argument of this section, and no element of it is a measurement of adoption, capability or maturity.

around 2020 and the attachment of tools to them around 2023 are three years and one enormous assurance question apart; the two right-hand boxes are a single generation in the argument below and a single row in Table 1.1, and they are drawn separately only to mark where reading turned into acting. Its axis, in turn, is proportional as far as 2020 and expanded thereafter, so that the last two milestones can be labelled apart; that is why no tick falls at the 2023 boundary the two right-hand boxes share, and why nothing should be inferred from the horizontal spacing of the final pair.

1.2.1 Rule-Based Consultation: Enumerable Behaviour, Unusable Workflow

The founding generation encoded expert knowledge as production rules with explicit uncertainty arithmetic and, importantly, an explanation facility that could recite the rule chain behind any conclusion [191]. It is easy to be condescending about certainty factors, and easy to forget how much these systems got right. Their behaviour was enumerable: the set of reachable conclusions was finite, inspectable, and testable, and every recommendation came with a derivation that a domain expert could audit line by line. Modern work on provenance and citation is in large part an attempt to recover a property that 1976 already had and that 2020 lost.

Their performance was also real. In the first blinded evaluation of a medical artificial-intelligence system, infectious-disease faculty judged the program's antimicrobial recommendations at a level comparable to the human experts, without knowing which recommendations were machine-generated [240]. That result should be sobering for anyone who believes expert-level accuracy is sufficient for adoption, because these systems never left the research setting. They failed on three counts that have nothing to do with accuracy. Knowledge acquisition did not scale: every rule was hand-authored, and maintenance grew superlinearly with coverage. Scope was brittle at the edges, since a consultation system for bacteraemia has no graceful behaviour when the patient's real problem is something else. And the interaction model was wrong: the clinician had to stop, launch a separate program, and type in structured findings that the record already contained. The generation's own leading figure

articulated the correct framing at the time — decision support as a tool for a clinician who retains ultimate responsibility for diagnosis and therapy [190] — and that framing, not the rule engine, is the durable contribution.

1.2.2 Order-Entry Alerting: Real Benefit, and the Birth of Override Culture

The second generation abandoned consultation for interception. Rather than asking clinicians to seek advice, it embedded checks in the path of an action they already had to take, namely writing an order. This is the only generation with an unambiguous positive effect on serious harm: a controlled study of computerised order entry with decision support and a team intervention showed a substantial reduction in serious medication errors [16], and the finding is durable enough that order-entry alerting became a *de facto* standard of care. The mechanism is worth naming, because this book reuses it: the system was placed at a choke point through which the consequential action had to pass. That is complete mediation *avant la lettre*, and it reappears as Principle 6.2 in Chapter 6.

The costs arrived quickly and were of two kinds. The first was technology-specific new error pathways. Detailed sociotechnical study of a widely deployed order-entry system documented fragmented displays, misleading dose-selection defaults, and inventory-driven rather than clinically driven presentation, all of which generated errors that had not existed before the system [105]. The second was worse, because it involved mortality: a rapid commercial implementation at a children's hospital was associated with an increase in mortality, an outcome its authors attributed less to the software's logic than to the reorganisation of work around it, including delays in obtaining medications and the loss of face-to-face coordination [77]. That study should be cited with its disputes attached. It is a single-centre, uncontrolled before-and-after comparison at one institution; its causal reading was contested at the time on grounds of case mix, secular trend and concurrent transport-process changes, and implementations of the same product elsewhere reported no comparable increase. The defensible reading is therefore narrower than the headline and more useful: it is evidence that the manner

of introduction can dominate the content of the logic, not evidence that order entry raises mortality. Nothing in that story requires a machine-learning model. It requires only that a system be inserted into a workflow at scale without staged rollout and monitoring, which is precisely the failure mode that Chapter 11 is written to prevent.

The lasting pathology of this generation, however, is alert fatigue. When override rates were measured in the meaningful-use era, roughly 93% of drug interaction alerts were overridden overall and about 95.1% of drug-drug alerts, with no evidence of improvement over time and no support for the comfortable theory that low alert volume explains high override rates [24]. An override rate above nine in ten is not a user-education problem. It is a specificity failure that has trained an entire profession to dismiss machine advice reflexively, and any agentic system that speaks to clinicians inherits that trained response as its prior. The meta-regression evidence points the other way as well: among features associated with effective decision support, requiring the clinician to supply a reason for overriding advice stood out strongly [176]. The lesson is not that friction is good. It is that attention spent must be matched by information gained, a budget made explicit in Chapter 8.

1.2.3 Statistical and Deep-Learning Support: Strong Internally, Fragile Externally

The third generation replaced hand-authored rules with functions fitted to data, which solved the knowledge-acquisition bottleneck and introduced a new one. Internal validation results were and are excellent; external validity is where the generation founders. The canonical demonstration is the external evaluation of a widely deployed proprietary sepsis prediction model, which performed substantially worse in local hospitalised patients than the vendor's figures implied, while generating a considerable alert burden [229]. The point is not that one vendor was careless. The point is structural: a model fitted to one site's population, documentation habits, and care processes encodes those habits, and the encoding is invisible in the model's stated intended use.

This generation also made the field's equity problems concrete. When the outcome a model predicts is a proxy for the outcome that matters — health care cost standing in for illness burden — the resulting risk scores can be systematically miscalibrated across racial groups even when race is not an input [149]. That finding matters for a book about retrieval agents because the mechanism is not confined to models. A retrieval pipeline that ranks by citation frequency, or by which service authored a note, or by which patients have complete data, is choosing a proxy target in exactly the same way and with exactly the same exposure.

1.2.4 What the Evidence Base Actually Says

Two claims are routinely conflated in the literature and in vendor materials, and separating them is the most useful thing this section can do. The first claim is that decision support changes what clinicians do. The evidence for that is reasonably strong. The second is that decision support changes what happens to patients. The evidence for that is weak, and it is weak for interesting reasons.

The definitive early synthesis found that decision support systems improved practitioner performance in the substantial majority of trials that measured it, while patient outcomes were reported far less often and improved rarely — and, critically, that effects were larger in evaluations conducted by the system's own developers [67]. Fifteen years later, a large-scale meta-analysis with absolute rather than relative effect measures reached a compatible and sharper conclusion: computerised decision support produces small-to-moderate absolute improvements in process measures and changes clinical endpoints little if at all [110]. That is the realistic prior against which any claim about an AI agent's clinical benefit should be read. Two mediating facts explain much of the gap. Uptake is dismal: a systematic review of provider use found that only about a third of available opportunities to use a system actually resulted in use, and that uptake is almost never reported in the first place [107]. And acceptance is not a fixed quantity: determinants shift over the deployment lifecycle, from expectations about outcomes early on to entrenched workarounds years later [145]. A system with perfect logic that is consulted in a third of eligible situations, and routed

around by year three, cannot move an endpoint no matter how good its recommendations are.

Table 1.1 summarises the comparison; the third column is where the discontinuity lies. The first three generations were bounded by construction: a rule engine can only conclude what its rules permit, an alert can only fire or not fire, and a risk model can only emit a number in a known range against a fixed input schema. A tool-using language model has no such intrinsic bound. Its output space is the set of strings, and with tools attached a subset of those strings are actions; nothing in its construction keeps a recommendation inside its validated scope, and nothing keeps it from calling a write endpoint. Bounds must therefore be imposed from outside, by an architecture — the entire argument of Chapter 6, and the reason this book treats autonomy as a per-action property rather than a system property.

Table 1.1. Four generations of clinical decision support. The right-most column is the operative one: each failure mode is inherited, not replaced, by its successor.

Generation	Encodes	Behaviour bounded by	Evidence / failure mode
Rule-based consultation (1970s–80s)	Hand-authored rules, uncertainty arithmetic, explanations	Finite, enumerable rule set, fixed at authoring time	Expert-level when blinded; never adopted: authoring did not scale, no workflow fit
Order-entry alerting (1990s–2000s)	Order-level constraints, dose checks, interaction tables	Deterministic interception at a mandatory action point	Fewer serious medication errors; alert fatigue, override above 90%, new error paths
Statistical and deep-learning models (2010s)	A fitted function from record features to a risk score	Fixed input schema and output range; locked version	Strong internal validation; weak external validity, drift, proxy-target bias
Retrieval-augmented agents (2020s–)	Parametric knowledge, a mutable corpus, and tools that read and write	Nothing intrinsic: bounds must come from an external control plane	Benchmark-level knowledge; unbounded output, ungrounded citation, unmediated action

Engineering Practice 1.1 Interrogating an effectiveness claim before a pilot

Vendor and preprint claims about clinical AI are almost always claims of the first kind — clinicians did something different — presented in the grammar of the second. Four questions separate them, and all four can be asked in a first meeting. *Which claim is this?* Ask for the primary endpoint as registered, and treat a process measure described in outcome language as a process measure. *Who evaluated it?* Developer-conducted evaluations reported larger effects two decades ago [67] and there is no reason to think the incentive has changed; an independent site, an independent analyst, or a pre-registered protocol is the minimum correction. *What was the denominator?* Uptake is the mediator that dissolves most reported gains, and it goes unreported in the great majority of studies [107], so ask what fraction of eligible situations resulted in use rather than what happened when the system was used. *Which generation's failure mode does it inherit?* Every row of Table 1.1 is inherited, not replaced: ask for the override rate if it alerts, the external-validation curve if it scores, and the unsupported-citation rate if it retrieves. A supplier who can answer all four is not thereby correct, but a supplier who cannot answer any of them is selling the narrative rather than the system.

1.3 Medical Foundation Models: Capability Without Accountability

The fourth generation arrived with a benchmark result and a narrative. The result was that a general-purpose language model performed at or near the passing threshold on United States medical licensing examination questions without domain-specific training [109]. The narrative, which the original authors were more careful about than their readers, was that the model therefore possessed clinical competence. Subsequent work substantiated the underlying capability claim in a more disciplined way: medical question-answering suites were assembled, and multi-axis human evaluation of long-form answers along factuality, potential harm, and bias

became the standard rather than accuracy alone [193]. Scores then climbed quickly, with the second generation of medically tuned models reaching around 86.5% on the USMLE-style MedQA set together with ensemble refinement and retrieval-conditioned grounding [194]. The result generalised beyond any single laboratory: a general-purpose model with careful prompting was shown to exceed the earlier medically fine-tuned systems on the same suites [147], a finding that its authors were themselves careful to frame as a statement about prompting rather than about medicine [148]; openly released medically adapted models reproduced much of the effect [35, 209]; and the multimodal successors extended it to images and long-form clinical dialogue [178]. The benchmarks themselves broadened in parallel, from single-source examination items to multi-subject question banks [156] and to diagnosis-oriented sets built from real presentations [92]. On the narrow question of whether a language model can encode a great deal of textbook medicine, the literature is settled: it can.

What that capability predicts about clinical work is a separate question, and here the evidence leans strongly negative. Examination items are epistemically generous in ways real encounters are not. They are complete: all the information needed is in the vignette. They are static: nothing changes between the question and the answer. They are single-turn: no follow-up question must be asked, no test ordered, no missing value chased. They are scored against a fixed key: exactly one option is correct, and abstention earns nothing. And they are drawn from the same distribution as the training corpus, since licensing questions and their explanations are abundantly represented on the public web. Strip those affordances away and performance does not degrade gracefully; it collapses. Recasting the same medical questions into a sequential, tool-using, patient-interactive setting — where the agent must elicit history, order and interpret investigations, and commit to a diagnosis under a query budget — produces accuracies far below the static-benchmark figures for the same underlying models [183]. The same pattern appears in agentic evaluation grounded in a virtual electronic record:

on realistic record-mediated tasks expressed against a standards-based interface, current agents are nowhere near the reliability their question-answering scores would suggest [90].

Failure Mode

Assuming benchmark performance transfers to the bedside.

The most common design error in clinical agent projects is to select a model on question-answering leaderboards and then treat the remaining work as integration. Three distinct gaps make that inference invalid. *Task form*: examination items supply complete information and reward guessing, whereas clinical work requires information-seeking and rewards well-calibrated silence. *Distribution*: licensing questions are near-training-set, whereas a local record's abbreviations, templates and formulary conventions are not. *Metric*: single-shot accuracy says nothing about consistency, and a system that is right 85% of the time on independent trials of the *same* request is not 85% safe — it is unpredictable, which is worse. Treat every benchmark number as an upper bound obtained under maximally favourable conditions, and require an evaluation on your own records, in the target interaction form, before any autonomy is granted. The harness in Appendix B exists for this purpose.

Two further findings should discipline expectations. First, the evaluation literature itself is thin where it matters most: a systematic review of testing and evaluation of health care applications of large language models found that only a small minority of studies used real patient data and that fewer still reported anything resembling deployment metrics [18]. We are therefore reasoning about clinical systems largely from evidence collected outside clinical settings. Second, and most instructive, capability does not automatically transfer to the human-machine team. In a randomised trial in which physicians were given access to a language model for diagnostic reasoning, the assisted physicians did not significantly outperform those using conventional resources — even though the model working alone scored higher than both physician arms [73]. Read that

result carefully, and resist reading more into it than it contains. It is not evidence that the model was weak. Nor, strictly, is it evidence about interface design: the trial manipulated access to the model, not the form in which its output was presented, so it establishes that capability did not transfer to the team and is silent on why. That the untouched variable — how a capable model’s output reaches a busy clinician, and what checking it costs her — is the one that governs the transfer is this book’s hypothesis about the null, not a finding of the study, and exercise 5 asks the reader to design the trial that would test it. The oversight plane of this book exists because that hypothesis is, on present evidence, the most plausible explanation available and because no deployment can wait for it to be settled.

Finally, capability arrives entangled with content the system should not reproduce. Commercial models have been shown to restate debunked race-based clinical claims and to answer inconsistently across repeated runs of the same prompt [151], and larger-scale audits show that triage and management recommendations shift with sociodemographic descriptors while the clinical content of the case is held constant [150]. A parametric model cannot be corrected by editing a source, because there is no source to edit. This is the accountability half of the section’s title: the fourth generation delivered a step change in capability with no corresponding mechanism for attribution, correction, or scope control.

1.4 Retrieval as the Pivot

Retrieval-augmented generation is best understood as an architectural response to that accountability gap rather than as an accuracy technique. The original formulation paired a parametric generator with a non-parametric index that could be inspected, swapped, or updated without retraining the generator [117]. Chapter 4 gives the formal treatment, including the definition of the retrieval-augmented pipeline itself (Definition 2.1), of a clinical claim (Definition 4.2), and of groundedness (Definition 4.3); the informal version suffices here. The system is given a question, it selects a set of passages from a governed corpus, and it is asked to answer using those passages and to say which passage supports which sentence.

The empirical case is real but modest and should be stated as such. Pooled evidence across biomedical studies gives an odds ratio of about 1.35 (95% confidence interval 1.19–1.53) for improved performance with retrieval augmentation across twenty studies [123]. Clinician-adjudicated evaluation of a retrieval system over curated clinical resources found improvements in factuality, completeness, and user preference relative to a standard model, along with better behaviour under adversarial prompting [242]. Those are worthwhile gains, not transformations, and the pattern of the gains matters more than their size: retrieval helps most on questions whose answers are stated somewhere in a document and helps least on questions requiring synthesis across a patient’s trajectory.

What retrieval genuinely buys can be reduced to three properties, and it is worth being precise about them because each maps to a distinct engineering obligation. **Currency** is the first: because the index is separable from the weights, a superseded guideline can be corrected by reindexing rather than retraining, and the correction propagates to the next query rather than to the next model release. **Attributability** is the second: the answer can carry pointers to the passages that licensed it, which converts an opaque assertion into a checkable one and, incidentally, is what regulators mean when they ask whether a clinician can independently review the basis of a recommendation. **Scope control** is the third and the most underrated: because the retriever decides what the generator may see, the corpus boundary becomes a policy instrument. Access control, patient consent scope, institutional formulary, and jurisdiction can all be enforced at retrieval time, which is also the only place they can be enforced cheaply.

Two properties retrieval does not buy, and conflating them with the first three is the origin of most disappointment in clinical RAG projects. It does not buy **corpus correctness**. If the indexed guideline is out of date, if a local protocol contradicts a national one, or if a note asserts a diagnosis that was later refuted, retrieval will faithfully surface the error and the generator will present it in confident prose with a citation attached. Guideline-grounded retrieval systems have been shown to still produce potentially harmful answers, and to degrade specifically on case-based rather than factual questions [133].

Nor does it buy **completeness**. Retrieval returns what matched a query; it cannot tell the reader what was absent from the corpus, absent from the record, or absent from the query. An answer about anticoagulation that never mentions the patient's scheduled procedure is not wrong about anything it says, and is dangerous nonetheless. Silence about omissions is the hardest failure to detect precisely because there is nothing to check a citation against.

A third caution belongs here rather than in Chapter 3, because it undermines the naive version of the attributability claim. Displayed citations are not verified citations. Automated assessment of how well model outputs are supported by the medical references they cite — including outputs from systems with live search — finds that a substantial share of statements are not supported by the source attached to them [232]. A citation is a hyperlink, and a hyperlink is a claim about support that must itself be checked. The provenance plane in Chapter 3 exists to make that check mechanical and its record tamper-evident.

1.5 From Retrieval to Agency: Tools, Plans, and Consequential Action

Retrieval alone leaves the system in a comfortable position: it reads, it writes prose, and a human decides. Agency dissolves that comfort. The pattern that made agency practical interleaves reasoning steps with tool calls, so that the model can decide to search, read a result, revise its plan, and search again, leaving an inspectable trajectory behind it [238]. Applied to clinical work, the tools are not toys. They are record queries, terminology lookups, dose calculators, order-set constructors, message composers, and scheduling endpoints, and standards-based record interfaces make all of them reachable through documented, authorised application programming interfaces rather than bespoke integrations [90]. Once such an interface is attached, the distance between a generated string and a clinical effect is measured in milliseconds.

Definition 1.1 (Clinical knowledge agent). A **clinical knowledge agent** is a software system that, given a clinical question or an event in a care workflow, autonomously plans and executes a sequence of

retrievals and tool invocations over governed clinical and biomedical knowledge sources in order to produce an assertion, a recommendation, or a proposed action, together with the provenance required to check it. Four properties distinguish such an agent from a retrieval pipeline with a chat interface: it selects its own next step rather than following a fixed pipeline; it holds state across steps and may revise earlier conclusions; it can invoke effectful operations, not only read-only ones; and its behaviour is therefore bounded only by an external control surface, never by its own construction. The per-action delegation levels formalised in Definition 2.2 in Chapter 2 are the vocabulary this book uses to describe how much of that boundary is opened, for which action, under which conditions.

The asymmetry at the heart of this book follows directly from the definition.

Principle 1.2 (*Reading is recoverable, acting is not*). An erroneous retrieval or an ungrounded sentence is a defect in a proposal, and proposals can be discarded at zero cost to the patient. An erroneous effectful call — an order placed, a message sent to a patient, a problem list entry written, an appointment cancelled — changes the state of the world and of the legal record, and its cost is bounded below by the cost of detecting it. Consequently the assurance budget of a clinical agent should be allocated not by how often a capability is used but by how irreversible its effects are, and every effectful path must be mediated by a component that can refuse.

This principle explains the two most common architectural mistakes. The first is to secure the model and leave the tools unguarded, on the theory that a well-aligned model will not misuse them. Indirect prompt injection makes that theory untenable: content retrieved from an untrusted source can carry instructions that the model follows, so any document the agent reads becomes part of its instruction stream [75]. In a clinical setting this is not hypothetical harm. Simulated injections delivered through an ordinary integration channel have been shown to induce commercial models to recommend contraindicated agents, including drugs contraindicated in pregnancy, across a large majority of simulated patient dialogues, with the harmful recommendation persisting over subsequent turns

[114]. The relevant question is therefore not whether the model can be tricked — it can — but what the write endpoint does when it is. Chapter 5 catalogues the attack surface and Chapter 6 specifies the mediation that answers it.

The second mistake is to treat autonomy as a system-level dial. Real clinical workflows contain actions whose reversibility differs by orders of magnitude within a single conversation: retrieving a guideline, drafting a note, proposing a dose change, transmitting a prescription. Granting one autonomy setting to that whole set means granting the setting appropriate to the most dangerous member, in which case the system is useless, or the setting appropriate to the least dangerous, in which case it is unsafe. The only coherent alternative is per-action delegation with per-action evidence, which is why Definition 2.2 is defined over actions rather than systems.

MediFlow Case Study 1.1 MEDIFLOW: an anticoagulation question at 02:10

Throughout this book we develop MEDIFLOW, a pedagogical reference architecture for a governed clinical knowledge agent. Every patient, note, value and event associated with it is synthetic; nothing in this book is derived from a real record.

The running scenario is deliberately mundane. A synthetic patient, an 81-year-old admitted overnight with a suspected pulmonary embolism, has an estimated creatinine clearance that has fallen from roughly 55 to roughly 38 mL/min over three days, a weight recorded twice with a 6 kg discrepancy, an active order for a direct oral anticoagulant at a standard dose, an over-the-counter interacting agent mentioned only in a triage free-text field, and a procedure provisionally booked for 08:00 that appears in the scheduling system but not in the plan of care. The covering resident asks: *should this dose be reduced tonight, and if so to what?*

Answering well requires six capabilities that map one-to-one onto the six planes of MEDIFLOW. The agent must locate and reconcile facts across heterogeneous stores, including the free-text mention (data and knowledge plane); rank a national guideline, the local protocol, and a product label by authority,

currency, and applicability to renal impairment (retrieval plane); compose those into a dose recommendation that distinguishes what the guideline states from what it infers (reasoning plane); recognise that a dose change is an effectful action requiring prescriber authorisation and refuse to place it itself (control plane); attach to every sentence the source that licenses it, including the timestamp of the creatinine it used (provenance plane); and decide whether this case warrants interrupting a clinician at 02:10 at all, or whether it should wait for the morning round (oversight plane). Chapter 10 walks the full trace end to end, and Appendix A gives the specification.

1.6 Why Healthcare Breaks the Standard Agent Playbook

The engineering culture around autonomous agents has developed a default playbook: give the model tools, let it plan, evaluate on a benchmark, iterate on prompts, and ship behind a feature flag. Every element of that playbook assumes properties that clinical environments do not have. Four breaks matter, and the rest of this book is organised around them.

Break one: ground truth is contested and time-varying. General-purpose agent evaluation rests on the existence of a verifier. A unit test passes or fails; a web task reaches the target page or does not. Clinical questions frequently have no such verifier. Guidelines from different professional bodies disagree, a local protocol may deliberately depart from a national recommendation for good reasons, evidence grades change with new trials, and for a large class of decisions the honest answer is a range with conditions attached. Worse, the truth is indexed by time in two distinct senses: the state of the patient changes, and the state of medical knowledge changes. A recommendation that was correct at 02:10 may be incorrect at 06:00 because a potassium value returned; a recommendation that was correct last year may be incorrect now because a guideline was revised. This has three consequences for engineering. Answers must be timestamped against both the patient state and the knowledge version they used, which is a provenance requirement (Chapter 3).

Source authority must be an explicit, auditable ranking input rather than an emergent property of embedding similarity (Chapter 4). And evaluation must accept a set of defensible answers rather than a key, which is why the harness in Appendix B scores groundedness and abstention behaviour alongside accuracy. A benchmark that rewards a confident guess over a well-warranted range is not merely uninformative here; it actively selects for the wrong behaviour.

Break two: the data terrain is hostile. Standard agent designs treat retrieved text as data. In healthcare, retrieved text is a mixture of data, instructions, and adversarial content, and the boundaries are not marked. A referral letter, a patient portal message, a scanned outside record, and a device report all enter the same context window, and any of them can contain a string that reads as an instruction [75]. The clinical demonstration that such injections translate into contraindicated recommendations [114] closes the argument that this is a theoretical concern. Hostility, moreover, is not always adversarial: the same mechanism fires on benign workflow artefacts, so an agent can be misled by content that no one authored with intent. Layered on top of this is the sedimentary structure of Sect. 1.1, where duplication inflates apparent consensus, and the legal terrain of protected health information, where the cost of retrieving the wrong document is not a wrong answer but a reportable disclosure. The architectural response is to treat the corpus as untrusted input, to separate the control path from the data path, and to make retrieval itself subject to authorisation — the subject of Chapter 5 and its threat catalogue Table 5.1, and of the closed-world default Principle 6.1 in Chapter 6.

Break three: harm is asymmetric, delayed, and hard to attribute. In most agent domains, failure is fast and legible: the code does not compile, the purchase fails, the user complains. Clinical harm is none of these. It is asymmetric, because the loss function is not symmetric around the truth — under-anticoagulating a patient with a pulmonary embolism and over-anticoagulating an octogenarian with a recent fall are both errors, but they are not the same error, and a system tuned to a symmetric metric will trade them off wrongly. It is delayed, because the bleed happens on day four, in a different unit,

possibly at a different institution, and the pathway from the agent's suggestion to the event passes through several human decisions. And it is diffuse in attribution, because when a clinician accepts a recommendation the resulting action is legally and professionally hers, which makes the counterfactual — what she would have done unaided — unobservable in the ordinary course of care. Bias makes the asymmetry worse in a specific direction: when management recommendations shift with sociodemographic descriptors while clinical content is held constant [150], the harm concentrates in the populations least able to contest it. Three engineering obligations follow. Severity, not frequency, must drive the assurance effort, which is the risk-management posture of Chapter 7. Every consequential recommendation must be reconstructible after the fact, bit for bit, which is what Theorem 3.6 in Chapter 3 makes possible. And surveillance must be continuous rather than pre-market, because the signal arrives late; Chapter 11 treats drift detection and incident response as ongoing obligations rather than closeout activities.

Break four: oversight is a scarce clinical resource. The final break is the one most consistently ignored, and it is the reason this book gives human oversight a plane of its own. Every proposal to keep a human in the loop spends a resource — clinician attention — that is already oversubscribed, that varies by hour and by unit, and that does not scale with the number of agent invocations. The alert-fatigue evidence already established what happens when this budget is overdrawn: override becomes reflexive at rates above nine in ten [24]. The human-factors evidence establishes something more troubling, namely that oversight is not merely finite but unreliable in a patterned way. A meta-analysis across 106 experiments found that human-AI combinations performed on average worse than the better of the two alone, with the shortfall appearing in decision tasks rather than in content-creation tasks [218]. That sample is largely non-clinical, and the step from its task taxonomy to the judgements a clinical agent supports is an inference this book draws rather than a decomposition the study reports; it is worth making the inference anyway, because the randomised diagnostic-reasoning trial discussed above shows the same shape inside medicine, where a capable model plus a physician underperformed the model alone

[73]. Reflexive dismissal and uncritical acceptance are the two failure modes, and a design that reduces one tends to increase the other.

Treating oversight as an engineered resource means, concretely, three things that Chapter 8 develops. Verification cost must be minimised rather than assumed away: an output that requires the clinician to open four documents to check one sentence has, in practice, no oversight at all, however many citations it displays. Interruptions must be budgeted and prioritised against expected harm avoided, so that the 02:10 question in the MEDIFLOW scenario is escalated only if waiting until the morning is worse than waking someone. And reliance itself must be measured, because appropriate reliance — following correct advice and rejecting incorrect advice — is the only outcome that matters and neither accuracy nor satisfaction measures it. There is also a longer-horizon consideration: the capacity to oversee may itself erode with exposure, which means oversight capacity must be monitored as a system property and not assumed constant. Regulators have arrived at the same place from a different direction, requiring that overseers of high-risk systems be able to disregard, override, or stop the system and that they remain aware of automation bias [58]; Chapter 7 maps that obligation onto implementable controls.

Taken together the four breaks explain the chapter's central and slightly counterintuitive claim. Agentic retrieval is a genuine discontinuity, because for the first time a system can perform the assembly step that Sect. 1.1 identified as the binding constraint — gather the fragments, reconcile them, and present a warranted answer at the moment of decision. But the assurance burden goes up rather than down, because the same step requires broad read access to privileged data, admits untrusted content into the reasoning path, produces outputs whose errors are fluent and plausible, and terminates in tools that change the world. Earlier generations were safe partly because they were weak. This one is not, and so its safety must be constructed.

1.7 What This Book Demands of a Clinical Agent

The normative position of this book takes two clauses, because presence and quality are different demands. A clinical knowledge

agent is defensible only if six architectural planes — ingestion, retrieval, reasoning and planning, control, provenance, and oversight — are engineered deliberately rather than inherited by accident; and a system that cannot exhibit a retrieval, provenance, control, or oversight plane *at all* should not be granted autonomy over any consequential action, regardless of how well it performs on any benchmark. The asymmetry between the six and the four is deliberate, and it is worth stating why, since a normative criterion that names the wrong set will be quoted against the wrong systems. Every system that answers clinical questions has an ingestion path and a reasoning step whether or not anyone designed them, so for those two the reviewer's question is how well and on what contract, not whether; their characteristic defects — metadata lost at chunking time, a plan step that influenced the answer and was never recorded — degrade the assurance the other four planes can offer rather than removing it outright. The remaining four are disqualifying by absence. Without a retrieval plane the system cannot say what its answer rests on; without a provenance plane it cannot show, afterwards, what it rested on then; without a control plane it cannot refuse; without an oversight plane there is no mechanism by which the human who remains accountable can correct it. Each of those absences removes the evidence on which a permission would have to be argued, which is why they disqualify rather than merely worry. Exercise 6 asks the reader to write the single failing test for each of the four. None of the six is a layer in a stack diagram; they are distinct sets of design obligations with distinct evidence requirements, and Chapter 2 arranges all six into the reference architecture (Sect. 2.4) that the remainder of the book instantiates.

The ingestion plane must turn heterogeneous sources — a guideline in a portable document, a note in a record, a product label, an outside document arriving through an exchange — into addressable units that still know where they came from. The obligation is a written contract per source class: what the unit of retrieval is, what identifies it stably, over what interval it is valid, what authority tier it carries, what access-control label it inherits, and how it was derived from the source. The failure is silent, which is what makes it dangerous: a chunk that no longer knows its guideline version, or

a fragment that no longer knows a consulting service wrote it four days ago, degrades every currency and authority decision made above it without ever raising an error. Given the sedimentary record of Sect. 1.1, this is also the plane at which duplication is either recorded as duplication or laundered into apparent corroboration. Section 2.5 gives its four source contracts.

The retrieval plane must treat access control, temporality, and source authority as first-class ranking concerns rather than filters applied after the fact. A retriever that ranks by semantic similarity and then removes documents the requester may not see has already leaked information through the ranking; a retriever that cannot distinguish a superseded guideline from a current one has no notion of currency; a retriever that treats a discussion forum and a national guideline as interchangeable evidence has no notion of authority. Chapter 4 builds the plane, defines groundedness and abstention formally (Definition 4.3, Definition 4.6), and is candid that the field lacks a settled evaluation methodology for any of this.

The reasoning and planning plane chooses the next step, carries state across steps, and assembles the candidate answer or payload. Its obligation in a clinical setting is not cleverness but legibility: the trajectory — the intent behind each tool call, the result returned, the claims the answer commits to — must be emitted as data rather than held in the model's context and discarded, because a plan step that influenced an answer and was never recorded turns every subsequent investigation into speculation [238]. This is where most engineering effort in the field currently goes and where the weakest safety arguments are made, since a critique generated by the same model that produced the error is not an independent check; Chapter 9 takes that question seriously enough to give it a theorem.

The control plane must completely mediate the path from model output to clinical effect, and it must degrade fail-closed. Complete mediation (Principle 6.2) means there is no path from a generated string to an effectful call that bypasses policy evaluation, including paths through retries, error handlers, and helpful convenience wrappers. Fail-closed means that when the policy engine, the provenance

service, or the terminology service is unavailable, the action is refused rather than permitted. Both are elaborated in Chapter 6; both are unglamorous; both are where real systems fail.

The provenance plane must make every assertion traceable to the evidence that licensed it and make the trace tamper-evident, so that a party holding an earlier checkpoint can detect retrospective editing. This is stronger than displaying citations, for the reason established in Sect. 1.4: displayed citations are frequently unsupported [232]. Chapter 3 formalises the provenance record (Definition 3.1) and gives the logarithmic inclusion proof (Theorem 3.7) and tamper-evidence result (Theorem 3.8) that make an audit affordable rather than aspirational.

The oversight plane must treat human attention as a measurable, budgeted resource, with explicit escalation criteria, measured reliance, and verification affordances designed to make checking cheap. Chapter 8 develops the plane, Chapter 9 asks how far machine reviewers can substitute for scarce human ones and reaches a deliberately cautious conclusion, and Chapter 11 closes the loop with telemetry, drift monitoring, and incident response. Chapter 10 shows all six planes operating on the MEDiFLOW scenario introduced above, Chapter 12 asks the reader to design, attack, and defend such a system, and six appendices supply the specification (Appendix A), the evaluation harness (Appendix B), the red-team playbook (Appendix C), the assurance dossier template (Appendix D), the glossary (Appendix E), and worked notes on selected review questions (Appendix F).

One clarification prevents a misreading of this programme. The argument is not that clinical agents should be denied the ability to act. A system that can only produce prose recommendations pushes the entire assembly and verification burden back onto the clinician and will be routed around, exactly as a third of the earlier generation's opportunities were routed around [107]. The argument is that the ability to act must be granted per action, on evidence, with a mediating component that can refuse, and with a provenance record sufficient to reconstruct what happened. Autonomy so granted is

an engineering property with a specification. Autonomy claimed at the level of a system is a marketing property with a legal exposure.

Key Takeaways

- ▶ The binding constraint in clinical knowledge work is the cost of assembling the right fragment about the right patient at the moment of decision, not the availability of knowledge. Note bloat, copy-forward, and fragmentation make the record a sedimentary deposit in which redundancy is mistaken for corroboration.
- ▶ Each generation of decision support attacked a narrower bottleneck than it claimed. Rule-based systems achieved expert-level accuracy and no adoption; order-entry alerting reduced serious medication errors and created override cultures with rates above nine in ten; fitted models solved knowledge acquisition and failed on external validity.
- ▶ The evidence base separates cleanly into two claims. Decision support changes practitioner behaviour: reasonably well supported. It changes patient outcomes: weakly supported, with small process gains and near-null endpoint gains, mediated by very low uptake.
- ▶ Examination-style benchmark performance is an upper bound obtained under maximally favourable conditions. Sequential, tool-using, patient-interactive settings degrade it sharply, and most published evaluation does not use real patient data.
- ▶ Retrieval buys currency, attributability, and scope control. It does not buy corpus correctness or completeness, and displayed citations are not verified citations.
- ▶ Reading is recoverable; acting is not. Assurance effort should be allocated by irreversibility, and every effectful path must pass through a component that can refuse.
- ▶ Healthcare breaks the standard agent playbook in four ways: contested and time-varying ground truth, a hostile data terrain, asymmetric and delayed harm, and oversight as a scarce clinical resource. The engineering response is

six deliberately engineered planes — ingestion, retrieval, reasoning and planning, control, provenance, oversight — of which four carry the disqualifying criterion: a system that cannot exhibit a retrieval, provenance, control, or oversight plane at all should not act consequentially, whatever its benchmark scores. Ingestion and reasoning are judged on how well they are built, not on whether they exist.

- Autonomy is a per-action property, never a system property.

Review Questions

1. Take a clinical question you can specify precisely (for example, whether a synthetic patient's antibiotic requires renal dose adjustment). Enumerate every distinct data element required, and for each record the source system, the expected latency, and one plausible reason the value could be wrong. Estimate the assembly cost in minutes for an unaided clinician, and state which of your enumerated elements a similarity-based retriever would be likely to miss entirely.
2. Using Table 1.1, write a one-paragraph argument that alert fatigue is a specificity problem and a one-paragraph rebuttal that it is a workflow-placement problem. Then state which position the override-rate evidence [24] and the override-justification finding [176] jointly favour, and why.
3. Construct a synthetic note that exhibits three sedimentation artefacts: a copy-forward plan for a resolved problem, a templated normal review of systems, and a diagnosis asserted in one note and refuted in a later one. Describe how each artefact would distort a top- k retrieval over sentence-level chunks, and propose one indexing-time mitigation for each.
4. Read the reported design of any published clinical RAG system and classify each of its claimed benefits as currency, attributability, or scope control. Identify at least one claim that actually requires corpus correctness or completeness, and state what evidence would be needed to support it.
5. Design an experiment that would distinguish the hypothesis “the model is not good enough” from the hypothesis “the human-machine interface is not engineered” as explanations for a null result of the kind reported by Goh et al. [73]. Specify the arms, the primary outcome, and the measure of reliance you would use.

6. For each of the four planes to which Sect. 1.7 attaches its disqualifying criterion — retrieval, provenance, control, oversight — write the single test whose failure you would treat as disqualifying for production use. Each test must be executable by someone with API access and no access to training data. Then write the test you would apply to the ingestion and reasoning planes, and say why its failure warrants a finding rather than a refusal.
7. Take the MEDIFLOW anticoagulation scenario and label each of the following as recoverable or irrecoverable: retrieving the national guideline, drafting a note addendum, computing a creatinine clearance, proposing a dose reduction to the prescriber, transmitting a modified prescription, cancelling the 08:00 procedure. For each irrecoverable action, name the evidence you would require before permitting it without a human in the loop.
8. Write the abstention policy you would adopt for a clinical agent as a decision rule over retrieved evidence, and state its two failure modes: the cost of abstaining when the answer was available, and the cost of answering when it was not. Explain why an accuracy-only benchmark cannot evaluate your rule.
9. Estimate the oversight budget of a single overnight covering clinician in interruptions per hour, justify the estimate, and compute how many agent escalations per hour are affordable at that budget. State what the system must do with the escalations that exceed it.

Discussion and Design Prompts

- D1. If a retrieval-grounded agent displays the passages behind every sentence, has it satisfied the requirement that a clinician be able to independently review the basis of a recommendation — or has it merely relocated the verification cost onto the clinician? What would have to be true of the display for the first answer to be correct?
- D2. The evidence that human-AI teams often underperform the better party alone [218] can be read as an argument against human-in-the-loop design. Argue the opposite position: that it is an argument about which loop, at which moment, over which decision. Where does your argument stop being persuasive?
- D3. Suppose an agent's recommendation is correct according to the national guideline in the index and incorrect according to a local protocol that deliberately departs from it. Which should the retrieval plane rank higher, and who in the institution owns that decision? Sketch the governance artefact that would record it.

- D4. Earlier generations of decision support were partly safe because they were weak. Is there a defensible version of deliberate weakness for clinical agents — capability that is withheld by construction rather than by policy — and what would it cost in clinical usefulness?
- D5. Consider a five-year horizon in which model capability continues to improve on every benchmark you can name. Which of the four breaks in Sect. 1.6 does that improvement dissolve, and which does it leave untouched or worsen?

Further Reading

Readers who want the historical spine of this chapter in the original should begin with the MYCIN monograph [191] and the blinded expert evaluation that followed it [240], then read Shortliffe [190] for the earliest crisp statement of the oversight principle this book operationalises. For the effectiveness evidence, the pairing to read is Garg et al. [67] and Kwan et al. [110]: two decades apart, methodologically different, and in agreement that process measures move and endpoints largely do not. Add Kouri et al. [107] for the adoption gap that explains much of that divergence, and Koppel et al. [105] and Han et al. [77] for the sociotechnical case that a correct system can still cause harm. Anyone designing alerting logic should read Roshanov et al. [176] and Bryant et al. [24] back to back before writing a single rule.

On the modern side, Singhal et al. [193] and Singhal et al. [194] establish what medical language models know and introduce the multi-axis human evaluation that accuracy alone cannot replace, while Schmidgall et al. [183] and Jiang et al. [90] are the two most useful correctives, showing what happens to those numbers when the task becomes sequential, tool-using, and record-mediated. For retrieval, Lewis et al. [117] gives the original formulation and Liu et al. [123] the pooled clinical effect; read Masannek et al. [133] and Wu et al. [232] immediately afterwards, since they define the limits of the grounding claim more precisely than any enthusiastic review will. Greshake et al. [75] and Lee et al. [114] are the shortest path to understanding why a read-only threat model is inadequate once tools are attached, and Vaccaro et al. [218] together with Goh et al. [73] is the pairing that most reliably changes how engineers think about the human in the loop. Each of these threads is picked up formally in the chapter indicated in Sect. 1.7.