purposive
INTELLIGENCE



Vertical AI Wins

where it matters

a synthesis of global adoption of Vertical AI in the workplace

About this Report

This report is an evidence synthesis, not a performance review. It draws on peer-reviewed research, independent benchmarks, and primary regulatory sources to examine a single question: in regulated industries such as healthcare and financial services, what actually determines whether an AI system can be trusted and deployed?

Its argument is deliberately measured. Where frontier general-purpose models have closed the raw-accuracy gap, it says so. The case it makes is narrower and more durable — that the dimensions regulated industries are obliged to care about, namely traceability, control, data governance, and compliance, are met most naturally by intelligence engineered for its domain, and that in India these are increasingly the written requirements of law rather than matters of preference.

The report names no products and makes no comparative performance claims about any vendor's systems, including those of its publisher. It is intended to inform the judgment of those weighing AI in high-stakes settings, and to be read alongside the primary sources it cites

Table of CONTENTS

01

The Gap Between Knowing & Doing

Why passing the medical licensing exam doesn't mean an AI is ready to practise. The evidence shows general models stumble exactly where stakes are highest — on safety — and that examination scores are misleading proxies for real-world competence.

Generalist and Specialist : What the Evidence Actually Shows

02

The honest scoreboard across medicine and finance. Domain-tuned models win consistently on bounded, document-grounded work and lose on open-ended recall — and that dividing line turns out to be exactly the one that matters in a regulated setting.

03

The Dimensions That Decide Adoption

Once raw accuracy is roughly a draw, the real decision is made elsewhere. This section examines the four dimensions a regulated buyer is actually obliged to weigh: controllability, traceability, data governance, and compliance.

The Indian Regulatory Reality

Where the technical argument meets the statute book. A reading of India's data-protection law and its draft guidance for clinical software, and what each actually requires of a deployed AI system.

04

05

Intelligence Engineered for Its Domain

What the evidence and the law, taken together, point to — and the practical question any organisation should ask first when weighing AI in a high-stakes, regulated setting.

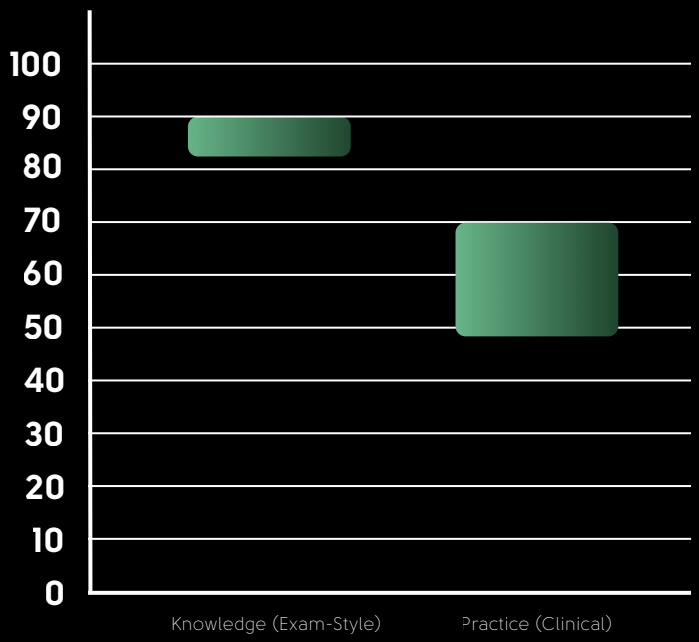
Act 1

The Gap Between Knowing & Doing

There is a comfortable story the technology industry likes to tell about artificial intelligence in regulated fields. It goes like this : the models have passed the exams. Large language models now score in the high eighties and low nineties on the United States Medical Licensing Examination and its equivalents around the world — matching, and in places exceeding, the performance of the physicians those exams are designed to certify. The newest reasoning models clear ninety-six percent on the question banks. If a machine can pass the test that gates entry to the profession, the reasoning goes, surely it is ready to do the work.

It is a tidy story. It is also, on the evidence, wrong — or at least dangerously incomplete. And the gap between that story and what the research actually shows is the reason this paper exists.

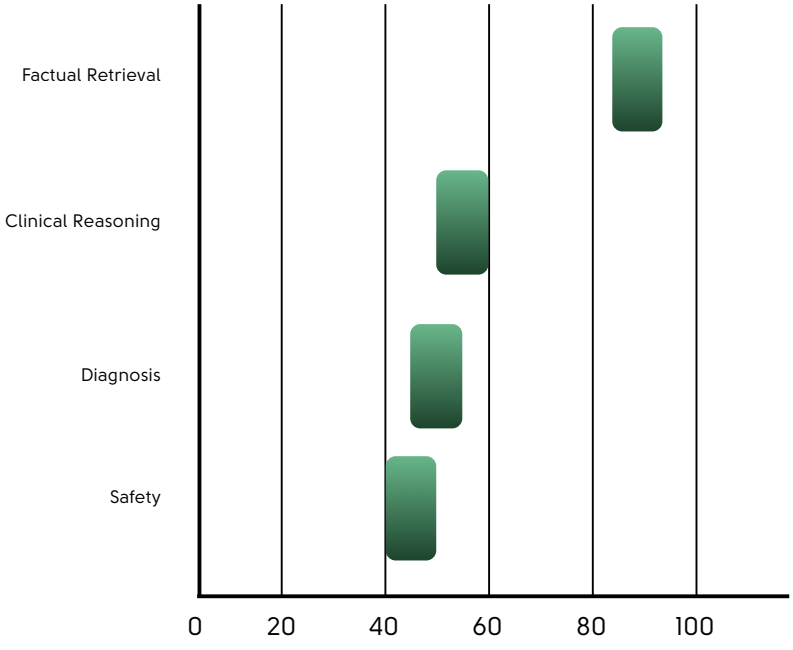
When a recent systematic review examined thirty-nine separate benchmarks built to evaluate



1.1 Knowledge v Practice

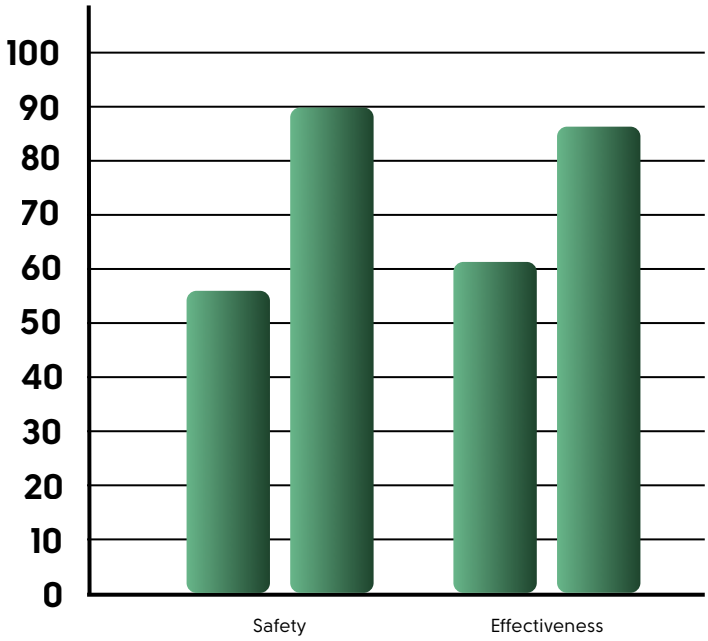
medical language models — collectively spanning over 2.3 million questions across forty-five languages and a hundred and seventy-two clinical specialties — it found something that should give any serious buyer of medical AI pause. The models' performance depended almost entirely on how they were asked. On knowledge-based tests, the multiple-choice format that mirrors licensing exams, leading models scored between eighty-four and ninety percent. But on the separate family of benchmarks designed to imitate actual clinical practice — open-ended diagnosis, reasoning through an incomplete patient record, managing a multi-turn consultation — performance fell into the forty-five to sixty-nine percent range. On the hardest diagnostic-reasoning benchmark, built from complex cases drawn from the world's top medical journals, the best reasoning model managed just under forty-six percent. The review is careful about what this comparison is and is not: the two sets of numbers come from different benchmarks, with different models and incompatible scoring methods, so they cannot be pooled into a single controlled measurement, and the authors hold only low certainty in direct cross-benchmark comparison. But the direction is not in doubt, and it is the direction that matters.language models are naturally good at — stays high, at eighty-five to ninety-three percent. But clinical reasoning drops to the fifties and sixties. Diagnosis falls further. And the most sobering finding concerns safety: on tasks that test whether a model can recognise a dangerous drug interaction, catch a contraindication, or avoid recommending harm, accuracy sat at just forty to fifty percent — the weakest category measured, despite being the one where errors are counted in lives rather than marks. A second study sharpens the point and, importantly, shows what begins to close the gap. A panel of thirty-two specialist physicians across twenty-six clinical departments built a benchmark expressly to test the two things that matter most at the bedside and that exams barely touch: safety and effectiveness. Two thousand and sixty-nine open-ended clinical scenarios, weighted so that life-threatening situations counted for more than routine ones. Six leading models were run through it. The average score was 57.2 percent — a moderate result that, again, looks nothing like the exam-room numbers. Safety, at 54.7 percent on average, consistently trailed effectiveness, at 62.3 percent. Every general-purpose model in the study displayed the same lopsided signature:

competent enough at producing a plausible answer, markedly weaker at ensuring that answer was safe. And when the scenarios were stratified by risk, every model dropped — a statistically significant decline of 13.3 percent — precisely when the stakes were highest. What makes that second study matter for the argument of this paper is the exception to the pattern. One model in the evaluation was not a general-purpose system but a domain-specific clinical model, purpose-built and tuned for medicine. It did not merely edge ahead; it scored 0.912 on safety and 0.861 on effectiveness, outperforming the next-best system by more than fifteen percentage points overall and by nearly twenty on safety alone. Its advantage was widest exactly where the general models were weakest — in high-risk specialties, with vulnerable patients, in the scenarios that carried fatal consequences. The study's own authors drew the obvious conclusion: that the model performed as well as it did not by accident of scale, but because it had been deliberately designed around the safety demands of the clinical setting. (It is worth noting, in the interest of rigour, that this particular model was evaluated by a team connected to its developer; the broader, vendor-neutral finding — that general models share a systemic safety deficit in specialised, high-stakes settings — does not depend on that single result and is the more important point.)



1.2 Performance by Task Type

That broader finding is not an isolated complaint. The same systematic review concluded, in its own words, that autonomous deployment of these systems is not currently justifiable, and that examination scores are insufficient and misleading proxies for clinical readiness — calling instead for validation against real practice and for human oversight kept firmly in the loop. The regulatory record bears the caution out: of more than a thousand AI- and machine-learning-enabled medical devices authorised by the United States Food and Drug Administration, not one built on a large language model had received clearance as of 2024. The gap between what these systems score and what they may be trusted to do is not merely an academic observation; it is already shaping who is allowed to deploy what.



1.3 Safety v Effectiveness - General v Specialist Models

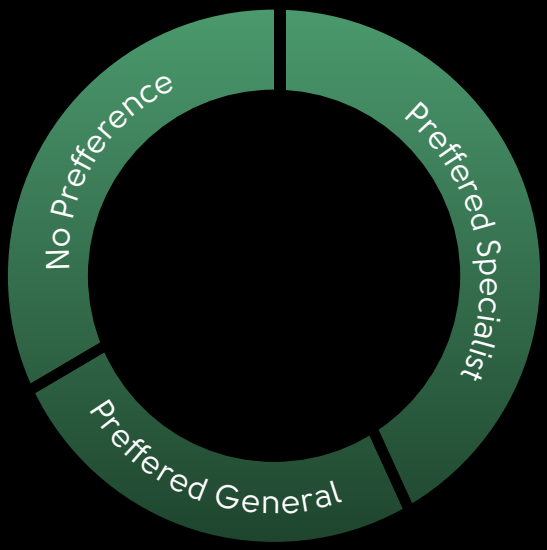
Two findings, then, sit at the foundation of everything that follows. First, examination performance is a misleading proxy for real-world competence in a regulated domain; the closer a benchmark gets to actual practice, the more general-purpose models struggle, and they struggle most on safety. Second, the early evidence suggests that the gap is not an immovable property of language models themselves but a function of how they are built and for what — domain-tuned systems begin to recover precisely the ground that generic ones lose.

This is the gap of the section's title, and it has a particular shape. It is not that AI cannot be useful in medicine, finance, or any other regulated field. It is that general-purpose AI, however impressive on a leaderboard, has not met the burden of proof that these fields demand — a burden measured not in average accuracy but in reliability under pressure, in the worst case rather than the typical one, in whether the system can be trusted when the consequences of being wrong are severe and irreversible. The chapters that follow examine what that burden actually consists of, what the evidence says about closing it, and why the answer points consistently toward intelligence engineered for its domain rather than borrowed from a general one.

Act 2

Generalist and Specialist : What the Evidence Actually Shows

The temptation, having established that general-purpose models stumble in clinical practice, is to reach for the obvious sequel: that domain-tuned models simply do it better. The evidence does support that conclusion — but only if you are careful about what "it" means. A domain-tuned model is a general model that has been taken further down a narrower road: trained again, or fine-tuned, on the texts and tasks of a single field until its instincts bend toward that field's way of using language. The honest picture from across medicine and finance is not that this process wins everywhere. It is that it wins consistently on a particular class of work, loses on another, and that the line between the two turns out to be exactly the line that matters in a regulated setting. Getting that distinction right is more useful than a clean scoreboard, and it is what the research actually shows.

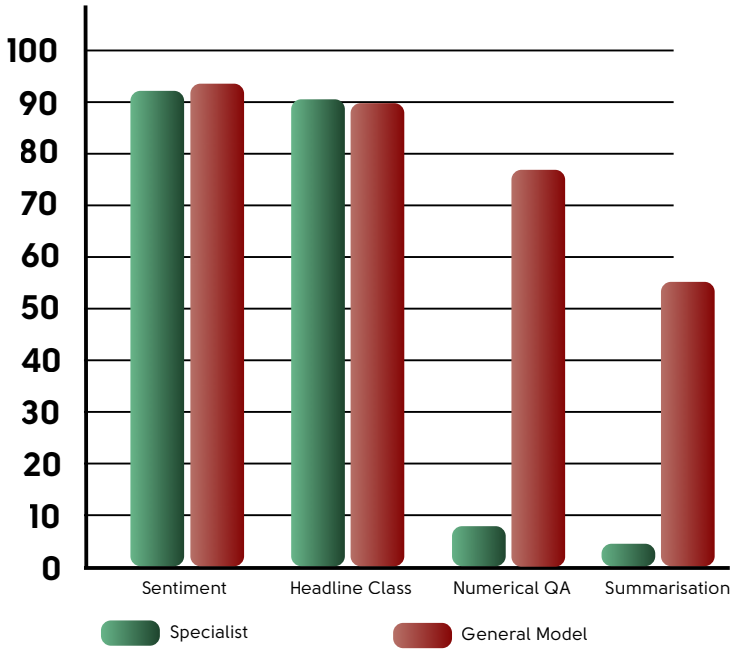


2.1 Clinical Documentation - Blind Physician Preference

Start with the clearest demonstration. In a blind, randomized study, practising physicians compared a general frontier model against a pair of healthcare-specific models built by fine-tuning on medical data. The blinding matters: the doctors judged each response on factuality, clinical relevance, and conciseness without knowing which system had produced it, so their preferences reflected the quality of the output rather than any reputation attached to the name. On the tasks that define real clinical documentation work — condensing a complex patient case into a usable summary, pulling the right answer out of a messy note — the doctors preferred the specialised model, and not narrowly. A model with eight billion parameters, small enough to run on a hospital's own hardware rather than in some distant data centre, was favoured over a vastly larger general model by roughly two to one on summarisation across all three dimensions. The point is not the size mismatch for its own sake, though it matters for cost and privacy; it is that the smaller, narrower system was better at the work precisely because it had been shaped around it.

The same study is honest about where that advantage stops. When the models were asked open-ended medical questions with no clinical context supplied — questions answered from broad recall rather than by reasoning over a specific document in front of them — the specialised model could not keep up with the generalist. The general model had been trained on an enormous and varied corpus, and that breadth gave it a reach the focused model simply did not have, even as the focused model held a narrow edge in factual accuracy within its own territory. This is the whole pattern in miniature: domain tuning buys depth on the bounded, context-rich tasks of a profession and trades away breadth on the open-ended ones. It is worth noticing that a regulated workflow — a discharge summary, a compliance check, an extraction from a filing — is built almost entirely from the first kind of task and very little from the second. Finance tells the same story with sharper numbers, and with the counter-evidence built in. An open, finance-tuned model evaluated across a battery of financial language tasks matched or beat a leading general model on sentiment analysis and headline classification — the routine sorting work of the field, deciding whether a statement reads as positive or negative, whether a headline concerns a particular event — scoring in the low nineties where the general model sat around ninety-three. But on the tasks that demand heavy numerical reasoning, the same specialised model collapsed: single-digit

accuracy on financial question-answering against the general model's seventy-six percent, and a near-total failure on financial summarisation. The lesson is not that domain tuning failed. It is that this particular model had been tuned for classification and never seriously trained on arithmetic reasoning or long-form synthesis, and it performed exactly as it was built to. Capability follows construction, not category — a model is good at what it was made to be good at, and the label on the box tells you almost nothing. What domain adaptation does buy, reliably, is sensitivity to the specialised meaning of language — the thing general models miss not because they are weak but because they are general. A separate finance study makes this concrete. It tested embedding models, which are systems that convert text into numerical representations so that passages with similar meaning sit close together — the machinery underneath search and retrieval. Across dozens of financial embedding tasks, the domain-adapted models outperformed their general counterparts, and, more tellingly, a model's score on the standard general benchmarks turned out to be a poor predictor of how it would do on the financial ones. A system can top the general leaderboards and still misread a field where a word like "liability" carries a weight of risk it never carries in ordinary speech. That same study supplies its own corrective against any triumphalism: on one kind of similarity task, a primitive method that does nothing more than count which words two passages share outperformed every sophisticated model tested. Specialisation is not magic, and newer is not automatically better; the right tool is an empirical question, not a branding one.



2.2 Finance - Where Vertical Wins and Where it Loses

The strongest caution comes from clinical AI itself, and it deserves to be stated plainly rather than buried. In a large benchmark drawn from real hospital records, the domain-specific medical models did not edge ahead of the general ones — they lost, and badly, with the leading general systems scoring in the high sixties on average while the best medical-specific model managed under forty. A naive reading of "vertical wins" would not survive contact with that result. But the reasons behind the failure are the resolution, not a refutation. Those medical models fell down on fundamentals that have little to do with medical knowledge: they could not reliably follow the instruction they were given, they buckled when fed the long, multi-page inputs that real records demand, and

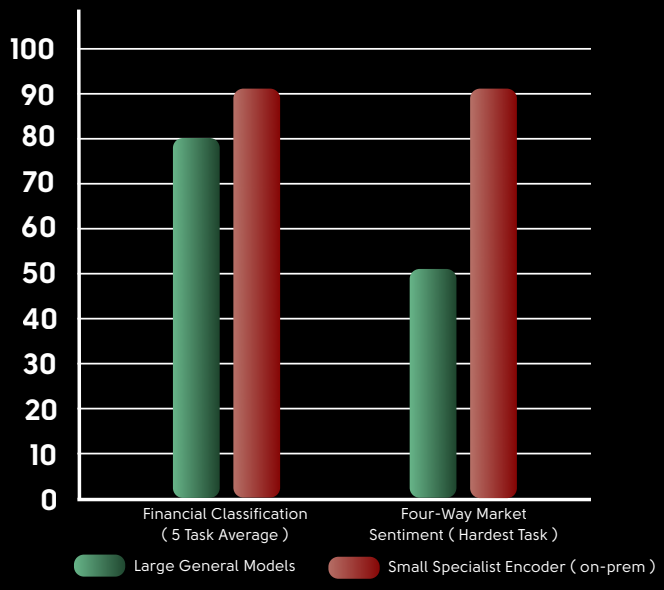
several could not consistently return an answer in the format the task required. They had been given a medical label without being given medical engineering. The finding does not unseat the thesis; it sharpens it. What separates the medical model that beats a frontier system from the one that loses to it is not the word "medical" but the quality and intent of the building underneath — the data it learned from, the tasks it was shaped for, the operational strain it was made to withstand. Assemble these results and a coherent picture emerges, one more defensible than any slogan.

On raw, open-ended reasoning and on broad recall, frontier general models are strong and frequently ahead; the gap that once flattered specialised systems on accuracy alone has largely closed there. On the bounded, context-dependent, document-grounded tasks that make up the actual working day of a clinician or an analyst, a well-built vertical model holds a real and repeatable advantage — and it tends to deliver that advantage at a fraction of the size, which is to say at a fraction of the running cost, and on hardware an institution can keep within its own walls. That last clause is where the argument has quietly been heading. Once raw accuracy is conceded to be roughly a draw on the tasks the leaderboards measure, the question of which system to trust inside a hospital or a bank stops being a question about scores at all. It becomes a question about everything the scores leave out — about where the data lives, whether an answer can be traced and checked, and who remains accountable when the system is wrong.

Act 3

The Dimensions That Decide Adoption

The previous section ended on a concession that turns out to be a beginning. Once raw accuracy on the open-ended tasks is granted to the frontier general models — and the honest reading of the evidence is that it largely should be — the question of which system a hospital or a bank ought to trust stops being a question about scores. The leaderboard has been conceded. What remains is everything the leaderboard was never built to measure: whether the institution can keep the system inside its own walls, whether an answer can be traced back to the document that produced it, and whether the data the system feeds on ever has to leave the building. These are not secondary considerations that a sufficiently accurate model would eventually render moot. In a regulated industry they are the considerations, and they decide adoption long before a single accuracy figure is consulted.

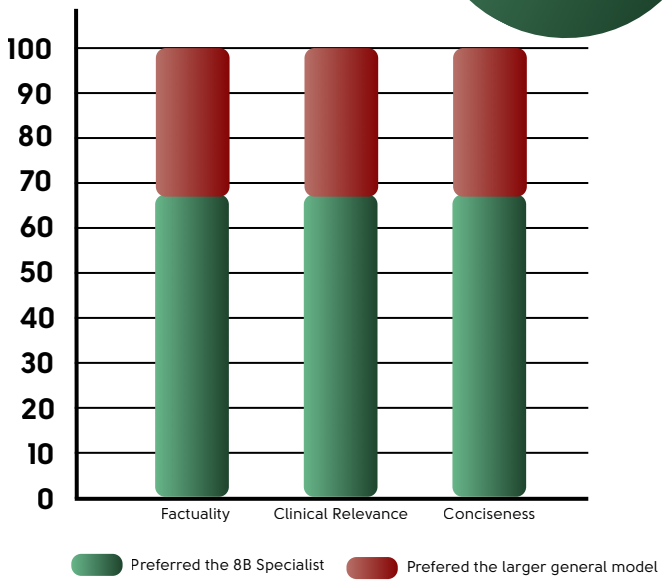


3.1 A small on-prem against large general models

Begin with the most concrete of the three, because it is the one the recent evidence demonstrates most plainly: control over where the model lives and what it costs to run. A specialised bidirectional encoder built for Chinese financial text — a small model, roughly a tenth of a billion parameters, the kind that runs on a single ordinary accelerator rather than a rack of them — was fine-tuned into a set of task-specific classifiers and set against the leading general models of the day on five financial classification tasks. The small specialist did not merely hold its own. It beat the large general systems by between roughly ten and twelve percentage points on average, and on the hardest of the tasks, a four-way market-sentiment judgement that turns on industry-specific conventions, the general models collapsed to below fifty-three percent while the specialist sat above ninety-two. The lesson is not that the small model is cleverer. It is that a model an institution can hold entirely in its own hands, retrain on its own data, and run at a fraction of the cost can also be the better performer on exactly the bounded, repetitive, high-volume work — labelling filings, sorting sentiment, tagging entities — that fills the regulated working day.

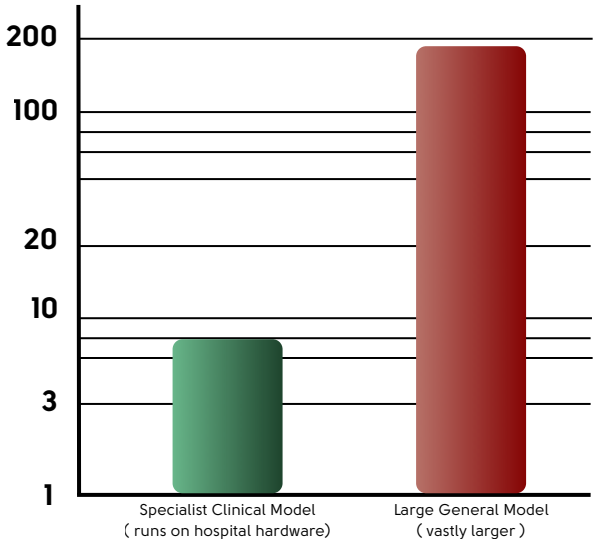
This echoes precisely what the clinical evidence showed earlier: the eight-billion-parameter medical model that beat a vastly larger general system on summarisation was favoured not in spite of being small enough to run on a hospital's own hardware, but in a way that made that smallness a feature, because cost and custody are not trade-offs against quality here — they often arrive together.That convergence matters because it reframes what "deployable" means. A closed, frontier model accessed over an API is, from the point of view of a regulated institution, a black box sitting on someone else's premises. It cannot be retrained on confidential internal data without sending that data outside; it cannot be frozen to a validated version while the vendor updates it underneath; and its behaviour cannot be fully audited because its weights and training data are not open to inspection. The finance survey examined in the previous section named this directly — that the strongest general model, for all its reasoning ability, remains closed and opaque, posing problems of data privacy, compliance, and customisability — and set against it the open, finance-tuned alternatives whose appeal is precisely that they offer control, on-premise deployment, and interpretability. The reason a regulated buyer reaches for the specialist is rarely that it answers a hard open-ended question better. It is that the specialist can be owned, and ownership is what governance requires.

The second dimension, traceability, is subtler, because it is not about the model generating text at all but about the machinery that decides what the model is allowed to see. Much of the real work in a regulated setting is not free generation but grounded retrieval — answering a question by first finding the right passage in the right filing, the right note in the right record, and reasoning only over that. The system that fetches those passages is an embedding model, the quiet infrastructure that converts text into numerical representations so that passages of similar meaning sit near one another, and it determines whether the answer an institution finally reads rests on the correct source or a plausible-looking wrong one. Here the domain-tuned systems again pull ahead: in the same finance work, retrievers built on the specialist encoder outperformed both the best open general embedders and a leading proprietary one across a battery of financial retrieval tasks, with the widest margins on exactly the document types — research reports, regulatory announcements — where pulling the wrong paragraph carries the most consequence. The earlier finance embedding study made the deeper point that gives this its force: a model's standing on the general retrieval leaderboards is a poor predictor of how it performs on financial retrieval, so an institution that selects its retrieval layer by general reputation may be installing a system that quietly misreads its domain. Traceability is not a feature one bolts on after the fact; it is a property of whether the retrieval layer was built to understand the language it is searching, and a generic one demonstrably is not.



3.2 Doctors preffered the smaller specialist

The third dimension, data governance, is where the first two meet and harden into a requirement rather than a preference. A model that runs on the institution's own hardware is a model whose training data and inference traffic never have to cross the boundary of the organisation — and in industries where the data is patient records or client financial positions, that boundary is not a convenience but the entire basis on which the data was lawfully collected. The finance survey was explicit that privacy concerns around sensitive financial data demand anonymisation and adherence to regulatory standards, and that the open, locally deployable models are the ones suited to use cases where interpretability and data confidentiality are paramount. The point generalises cleanly beyond finance: the same logic that keeps a bank's client data on the bank's servers keeps a hospital's records inside the hospital. A system whose custody can be guaranteed is one whose compliance posture can be reasoned about; a black box on a third party's infrastructure forecloses that reasoning before it begins.



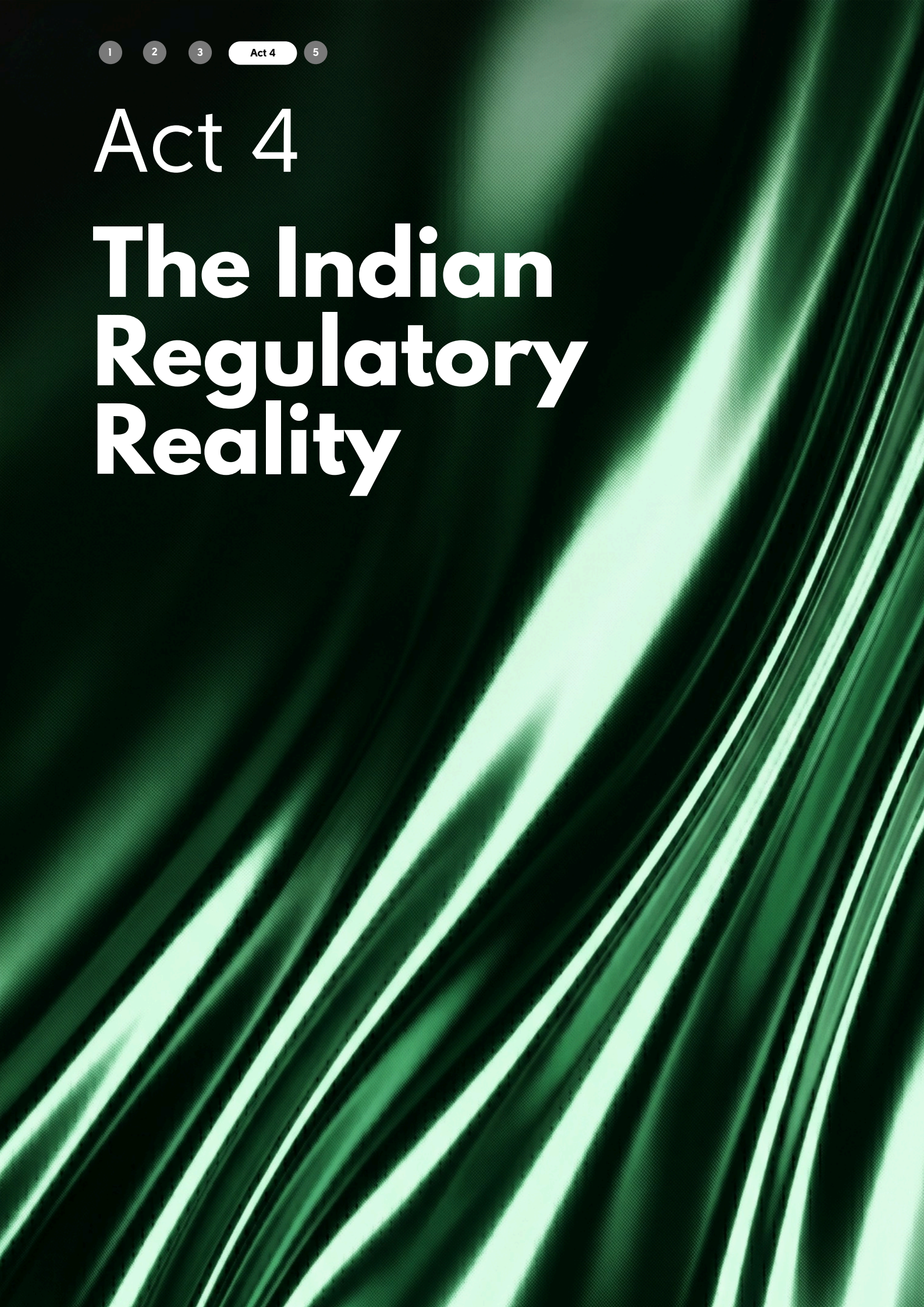
3.3 Custody and Cost : the size gap behind the performance

None of this should harden into the slogan the previous section warned against. The evidence for these dimensions carries its own honest qualifications, and they belong in plain view. The specialist's advantage over the general models in that classification study was not uniform: on the simpler task of recognising a person's name in a document the general models sometimes matched or beat the fine-tuned ones, and on the more open-ended task of pulling company names out of running text the general systems' broader generalisation occasionally told in their favour — the same depth-versus-breadth boundary that ran through all of Section 2, reappearing here. The authors of that work were candid, too, that their fine-tuned results retained a degree of run-to-run randomness, a reminder that a specialist is not a guarantee but a well-built tool whose reliability still has to be earned and verified. And the deployability argument cuts both ways: a small model an institution runs itself is a small model that institution must now maintain, validate, and take responsibility for, where the API at least outsources the operational burden. The case is not that the vertical system is unconditionally superior. It is that on the axes a regulated buyer is actually obliged to care about — custody, auditability, the physical location of the data — the vertical, controllable, locally deployable system answers questions the frontier general model cannot, and answers them by design rather than by accident.

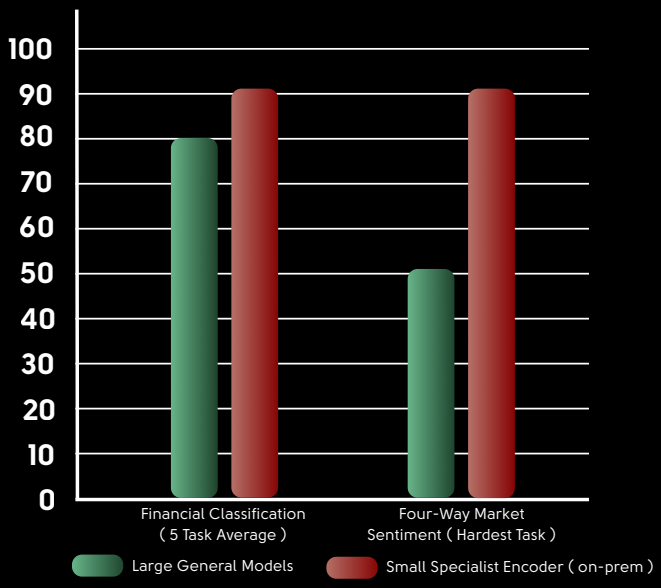
That is the shape of the decision as the technical evidence frames it. But "compliance" has so far been used loosely, as though it were a single property a system either has or lacks. In a regulated industry it is nothing of the kind. It is a specific, written body of law — about what may be done with personal data, about how a piece of clinical software must be evaluated before it touches a patient — and whether a system can be controlled, traced, and governed is ultimately a question asked and answered in those legal terms. What those terms actually say, in the jurisdiction this paper is written for, is where the argument turns next.

Act 4

The Indian Regulatory Reality



The previous section ended by arguing that compliance, custody, and auditability are not finally technical properties but legal ones — that whether a system can be trusted inside a hospital or a bank is a question answered in the language of statute. It is worth seeing what the statute actually says. India has, within a single recent stretch, moved two distinct bodies of law from aspiration to text: one governing personal data across every sector that touches it, and one governing the narrower case of software that makes clinical decisions. Read together, they describe with unusual precision the obligations a deployed AI system must meet — and in describing them, reveal which architectural choices are merely preferences and which have quietly become requirements.



3.1 A small on-prem against large general models

Begin with the broader. The Digital Personal Data Protection Act, 2023 — Act 22 of 2023 — is India's first comprehensive statute on personal data, and its arrival has been deliberate rather than sudden. Assented to in August 2023, it is being brought into force in stages: the definitions and the machinery of the new Data Protection Board took effect in November 2025, while the operative heart of the law — how data may be collected, the rights it grants individuals, the duties it imposes on those who process it — follows roughly eighteen months from that commencement, with one provision on children's data on a twelve-month track. The phasing means the substantive obligations are not hypothetical futures but dated certainties, their operational detail to be filled in by subordinate rules.

At its core the Act demands control and traceability over personal data — the same two properties the technical evidence kept pointing to, now rendered as enforceable duty. The law turns on a relationship between a "Data Principal," the individual the data is about, and a "Data Fiduciary," whoever decides how and why it is processed; the word fiduciary, with its freight of trust and accountability, is not accidental. A fiduciary may process personal data only for a lawful purpose, and only on one of two bases: the individual's consent, or a narrow set of "legitimate uses." Where consent is the ground, it must be free, specific, informed, unambiguous, and as easy to withdraw as to give, and each request for it must be preceded by a plain-language notice stating what data is wanted and why. An AI system that ingests personal data, in other words, cannot treat that data's provenance as someone else's problem; the lawfulness of every downstream use is tethered to a specific, recorded purpose at the point of collection. From there the Act reads almost as a design specification. A fiduciary must keep data accurate where it will inform a decision about someone, protect it with reasonable security safeguards, report any breach to both the Board and the affected individuals, and erase it once consent is withdrawn or the purpose served — carrying that erasure duty through to any processor acting on its behalf. The individual gains rights a system must be built to honour: to obtain a summary of what is held and a list of everyone it has been shared with, to demand correction or erasure, to seek redress, to nominate someone to act for her. Each is, in engineering terms, a requirement. The right to know what was shared presupposes that sharing was logged; the right to erasure presupposes the data can be found wherever it propagated. A system that cannot trace its own data lineage cannot satisfy them.The Act reserves its sharpest demands for two cases. For children's data it forbids tracking, behavioural monitoring, and targeted advertising aimed at children, and requires verifiable parental consent before processing at all. The second case is the "Significant Data Fiduciary," an entity the government may designate by the volume and sensitivity of the data it handles.

Such a fiduciary must appoint a Data Protection Officer based in India and answerable to its own board, commission an independent auditor to evaluate its compliance, and conduct periodic data-protection impact assessments — statutory obligations to make the system auditable by design and to locate accountability in a named person on Indian soil. That last point connects to the provision bearing most directly on where a system runs: the government may, by notification, restrict transfers of personal data to particular countries — a model that leaves cross-border flow open until specifically curtailed, but places data location squarely within the regulator's reach. The cost of error is not nominal. The Act's penalties run to two hundred and fifty crore rupees for failing to maintain reasonable security safeguards, and two hundred crore for failing to notify a breach — figures large enough to make governance a board-level question rather than a technical footnote. It would overstate the case to read the Act as a mandate for any particular kind of AI. It is technology-neutral, much of its operational detail awaits rules not yet fully populated, and a large general-purpose system run by a well-resourced provider can in principle meet every requirement. The honest claim is narrower and more durable: the Act converts a list of soft virtues — traceability, control, accountability, data locality — into hard obligations with named penalties, and a system built so that data stays in known places, every use traces to a recorded purpose, and a specific person can answer for it meets those obligations as a matter of structure rather than effort. Not that vertical systems are better, but that the properties they tend to have by design are the properties the statute now requires. If the DPDP Act follows personal data wherever it goes, the second body of law addresses a higher-stakes case: software that does not merely hold clinical data but acts on it. In October 2025 the Central Drugs Standard Control Organisation, the national medical-device regulator, circulated a draft Guidance Document on Medical Device Software under the Medical Devices Rules, 2017. Its draft status must be stated plainly, and the document states it of itself: issued for stakeholder comment, described as being for awareness rather than legal use, and pointing to the Drugs and Cosmetics Act of 1940 and the Medical Devices Rules as the binding instruments behind it. It is not yet law — but it is the clearest available statement of how the regulator intends to read existing law as it applies to clinical software. The guidance first draws a line much commentary blurs: not all clinical software is a regulated device. It separates software embedded in a device and driving its function — the firmware in a

A
LOW RISK
Informs clinical management in a non-serious situation.

B
LOW-MODERATE
Informs or drives management as the situation grows more serious.

C
MODERATE-HIGH
Drives clinical management, or acts in a serious situation.

D
HIGH RISK
Diagnoses or treats directly in a critical, time-sensitive situation.

pacemaker or insulin pump — from standalone "Software as a Medical Device" that performs a medical purpose in its own right, such as a tool analysing an image to triage a suspected lesion; and it sets aside hospital administration, scheduling, billing, and record-storage software entirely. The line identifies precisely the category domain-tuned clinical AI most often occupies — software that interprets clinical data to inform or drive a decision — and places it under the full weight of device regulation. For that category the draft sets a risk-based scheme. Borrowing the four-class structure of the Rules, it grades software from low-risk Class A to high-risk Class D by combining two judgments: the significance of the information provided — whether it informs clinical management, drives it, or is used to treat or diagnose directly — and the seriousness of the situation, from non-serious to critical. A tool diagnosing a critical, time-sensitive condition lands highest. Two rules sharpen the edge: where several could apply, the strictest governs; and software placed with non-specialist users in a serious situation may be elevated, on the reasoning that absent expert oversight the stakes rise. Risk here is a property of context and use, not of the algorithm alone. What the draft then demands of a higher-risk system is a documented account of its own trustworthiness — the auditability of the data-protection law, now extended to the decision-making software itself. A manufacturer must run a quality-management system across the full software lifecycle and assemble a technical file describing, among much else, the analysis methodology (rule-based or machine-learning, fixed or adaptive) and the software's degree of autonomy; document its architecture; validate it across the hardware and operating systems it will run on; demonstrate clinical evidence for its claims; and maintain version control and traceability sufficient for post-market tracking, so a fielded problem traces back to a specific build. Running through the risk-management expectations is attention to what makes software unlike a physical device: that it updates, deploys at scale instantly, can carry hidden bias, and can change behaviour with a change to its model.

That concern produces the draft's most forward-looking element. It contemplates an Algorithm Change Protocol — a pre-specified plan for how training data is managed, performance monitored, the model retrained, updates pushed, and a bad update rolled back — and treats post-market surveillance as continuing: a duty to watch for direct and indirect harm and security vulnerabilities, to handle recalls and field corrections, to halt or uninstall a fielded system, and to report serious adverse events within fifteen days. A model that learns and shifts after deployment is, in this framing, not a finished product but a system under continuous obligation, its behaviour observable, attributable, and reversible for as long as it is in clinical use. Here the resonance with the technical argument becomes explicit. To meet these expectations a clinical AI system must be controllable in the strict sense the regulator means — versions tracked, updates governed, outputs traceable to an accountable manufacturer, behaviour capable of being frozen or withdrawn. These are precisely the properties a contained, domain-specific, locally governable system offers by construction, and that a large general model accessed as an opaque external service — changing on a vendor's schedule, beyond the institution's control — struggles to guarantee. The point is not that a general model cannot be wrapped in such controls; it is that the draft asks for exactly the custody, traceability, and change-governance that domain-built systems treat as foundational rather than retrofitted.

Two laws, approached from opposite directions, arrive at a common destination. One reasons from the individual's data outward, the other from the clinical decision inward, but both end by demanding that an AI system be traceable in its handling of information, controllable in its behaviour, governed in its changes, and anchored to an accountable party within reach of the regulator. The draft guidance must be read for what it is — consultative, not yet binding, subject to revision — and the data-protection law for what it is, a technology-neutral statute no architecture is exempt from and none uniquely blessed by. Taken together, they make concrete what the earlier evidence could only assert: the dimensions on which domain-tuned systems hold their advantage are no longer matters of engineering taste. In the jurisdiction this paper is written for, they are increasingly the text of the law.



Act 5

Conclusion: Intelligence Engineered for Its Domain

The argument of this paper has moved in one direction throughout, and it is worth retracing the line before drawing it to its end. It began with a comfortable story — that AI which passes the licensing exam is ready to do the work — and showed that story to be, on the evidence, dangerously incomplete: the closer a benchmark comes to real practice, the more general-purpose models falter, and they falter most precisely on safety, in the high-stakes cases where being wrong is measured in lives rather than marks. It then asked whether domain-tuned systems do better, and gave the honest answer rather than the convenient one. They do not win everywhere. On open-ended reasoning and broad recall the frontier general models are strong and often ahead, and the accuracy gap that once flattered specialised systems has largely closed there. But on the bounded, document-grounded, context-dependent work that fills the actual day of a clinician or an analyst, a well-built vertical system holds a real and repeatable edge — and tends to deliver it at a fraction of the size, cost, and infrastructure.

That concession turned out to be the hinge of the whole argument. Once raw accuracy is granted to be roughly a draw on the tasks leaderboards measure, the question of which system to trust inside a hospital or a bank stops being a question about scores. It becomes a question about everything the scores leave out: whether the institution can keep the system within its own walls, whether an answer can be traced to the source that produced it, whether the data the system feeds on ever has to leave the building, and who remains accountable when the system is wrong. On those dimensions — control, traceability, governance, accountability — the advantage of a contained, domain-built system is not a matter of marginal performance but of architecture. A system that can be owned can be governed; a black box on a third party's infrastructure forecloses that reasoning before it begins. The paper's final move was to show that these dimensions are no longer a matter of engineering preference or institutional taste. In India they are increasingly the written text of the law. A comprehensive data-protection statute now converts traceability, control, accountability, and data locality into enforceable duties carrying penalties large enough to reach the board, and a draft guidance for clinical software — consultative still, but a clear signal of regulatory intent — asks of any system that informs a medical decision exactly the custody, version control, change-governance, and reversibility that domain-built systems treat as foundational. Two bodies of law, reasoning from opposite starting points, converge on the same requirement: that a deployed system be traceable, controllable, governed, and anchored to someone who can answer for it. The conclusion the evidence points to is therefore narrow, and deliberately so. It is not that vertical AI is smarter than general AI, nor that general models have no place in regulated work — the overclaim this paper has been at pains to avoid. It is that regulated industries are subject to demands that accuracy alone cannot satisfy, and that those demands are met most naturally by intelligence engineered for its domain: built around the safety profile of the field, deployable where the institution can keep watch over it, traceable in its outputs, and governable through its changes. The burden of proof these fields impose is not measured in average accuracy but in reliability under pressure, in the worst case rather than the typical one. Meeting that burden is less a question of how capable a model is than of how it was built, for what, and under whose control.

That is the practical lesson for any organisation weighing AI in a regulated setting. The first question is not which model scores highest, but which the institution can own, audit, contain, and answer for — because that, increasingly, is the question the law asks too. The evidence and the statute now point the same way. Intelligence that wins where it matters is intelligence designed for the place it has to work.

Explore what Vertical AI can do for you.

Get in Touch