



Peregrine.[Reason](#)

Peregrine.Reason

Technical Whitepaper

Clinical Reasoning Through Structured Disagreement

Version	0.2
Date	September 2026
Supersedes	v0.1, April 2026
Document type	Public technical and clinical architecture overview
Status	Research and development whitepaper
Phase I domain	Adult differentiated thyroid carcinoma — multidisciplinary team (MDT) reasoning
Development base	Andorra
Published by	The Peregrine.Reason programme, Andorra

Disclaimer

Peregrine.Reason is a research and development architecture for clinical reasoning support.

It is not an approved medical device, diagnostic system, autonomous treatment recommendation system, autonomous tumour board, or replacement for clinician judgment. It does not prescribe, calculate doses, generate orders, interpret images, or produce patient-facing output.

Nothing in this document is a claim of demonstrated clinical performance. The capabilities described are design requirements and research objectives. Whether the architecture delivers measurable value over simpler alternatives is an open experimental question, and the programme is deliberately designed so that the answer can be negative.

Any clinical deployment would require appropriate validation, clinical governance, data protection assessment, cybersecurity review, human oversight, and a documented regulatory assessment against the intended purpose and target jurisdiction.

1. Publication status and what this document is

This whitepaper describes the proposed architecture and development direction of Peregrine.Reason, and the evidence the programme intends to generate about it.

It is written to a single standard: claims are kept proportionate to the evidence that exists today. Where a component is specified but not built, it is described as a build requirement. Where a property is intended but not measured, it is described as a research question with a named test. Where a design choice may turn out to be unnecessary, that is stated rather than concealed.

Peregrine.Reason is currently positioned as a research and development system for clinical reasoning support. It is not presented as an approved diagnostic device, autonomous treatment recommendation engine, autonomous tumour board, or replacement for clinician judgment.

Phase I is a self-contained twelve-month development and formative-evaluation programme. Later product maturation, independent academic evaluation and any regulated market access are separate, later-gated activities.

Two horizons are described in this document, and they are kept apart deliberately.

Sections 3 to 17 describe the Phase I architecture and the evidence the funded programme will generate about it. Sections 18.1 and 18.2 describe long-term direction beyond Phase I. Long-term material is labelled as such wherever it appears, is not a Phase I deliverable, is not costed or committed in the Phase I programme, and is not offered as evidence that Phase I will succeed. A reader comparing this whitepaper against the Phase I programme documentation should find the funded scope smaller than the direction — that is intended, not an inconsistency.

2. Executive summary

Peregrine.Reason is a multi-model clinical reasoning architecture designed to support complex medical decision-making by making evidence, uncertainty, competing management pathways and minority safety concerns inspectable, rather than collapsing them into a single fluent answer.

The principal research object is the **Peregrine Moderator Engine (PME)**. PME is not a majority-vote mechanism and not a single-model summariser. It is a claim-level adjudication pipeline that:

- decomposes independent model contributions into atomic claims with provenance;
- links each material claim to case facts and versioned external evidence;
- constructs a typed argument graph of support, contradiction, qualification and dependency;
- classifies disagreement into clinically meaningful categories;
- builds outcome-separated benefit–harm objects with null-safe arithmetic;

- identifies decision pivots — the facts, thresholds or preferences that would change the balance;
- escalates minority safety concerns on consequence, not frequency;
- routes bounded, targeted challenges and records why positions changed or persisted;
- submits the result to a sealed two-pass independent safety review;
- and terminates in one of five explicit decision states, including *insufficient information*.

The clinician remains the decision-maker throughout. The output is an auditable decision map, not an instruction.

The failure mode the architecture is built against is not disagreement. It is **false consensus** — uncertainty that is still present but has become invisible, whether because a human meeting moved too fast or because a model panel shares a training lineage. Open disagreement can be examined. Invisible uncertainty cannot.

Two design commitments distinguish this from generic multi-agent debate:

No claim gains authority merely because more agents repeat it. Support counts do not select a pathway, do not override a safety concern, and do not appear as majority badges on the primary clinician view. They remain available as secondary audit detail.

A pathway can only be presented as preferred under an explicit, recorded rule. PME does not derive a winner by summing ordinal scores. Preference requires either that a single option survives admissibility testing, or that a documented dominance test is satisfied across pre-specified material outcomes. Where neither holds, the system returns structured equipoise, a conditional map, or insufficient information — and these are correct outputs, not failures.

Adult differentiated thyroid cancer is the Phase I formative-evaluation domain. The reusable reasoning architecture, not a single disease application, is the product objective.

3. Conceptual foundation: consensus is not truth

Agreement is informative. It is not a guarantee of correctness. Model agreement may reflect sound evidence — or shared training data, overlapping alignment methodology, anchoring on an early position, or repeated retrieval of the same source. In a panel of language models, agreement is partly a measure of how similar the models are.

The symmetric point matters just as much, and it is where systems of this kind most often fail:

Disagreement is not intrinsically valuable either. A system rewarded for finding disagreement will manufacture it. Vocabulary differences, restatements, hedging and immaterial divergence must not be promoted into a clinical brief. Alert flooding destroys a decision-support tool as effectively as false certainty does.

Peregrine. Reason therefore preserves disagreement **only where its facts, evidence, clinical consequences or patient-value implications are material**. Separating material dissent from noise is a core engineering problem, not a framing device, and dissent-flag precision and recall are explicit evaluation endpoints (Section 12).

3.1 From one doctor, to the multidisciplinary team, and back to one machine?

Modern medicine moved deliberately away from individual expert opinion toward multidisciplinary decision-making. That was progress, and the reason is instructive.

A single clinician, however experienced, has a characteristic field of view. A surgeon may reason about whether a lesion can be removed; a generalist asks whether it should be. A radiologist identifies an imaging pattern that changes operative risk. A pathologist is appropriately cautious about a borderline diagnosis. An oncologist reasons in probabilities and long horizons. A specialist nurse understands the patient's resilience, support structure and actual priorities.

The multidisciplinary team exists because complex problems are better approached when several independent perspectives examine the same question from different angles. Medicine moved from centralised reasoning toward distributed reasoning — and it only works when disagreement is allowed to surface. The value of an MDT is not that everyone agrees. It is that assumptions are challenged before irreversible decisions are made.

If medical AI simply carries the sequence one step further —

one doctor → multidisciplinary team → one model

— that is not progress. It is regression presented as innovation.

A single large language model may sound balanced and authoritative. It may summarise guidance well and answer in seconds. But if it collapses uncertainty into one confident response, it reproduces precisely the weakness that multidisciplinary medicine was built to overcome. Medicine moved beyond the single expert for a reason, and an architecture that restores the single expert in silicon has not learned that reason.

3.2 Why false consensus is the greater danger

In high-stakes environments the greater danger is usually not open disagreement. Open disagreement can be examined, tested and resolved.

The greater danger is **false consensus**: uncertainty that is still present but has become invisible.

In human teams, false consensus arises when discussion moves too quickly, when hierarchy quietly silences doubt, when guidelines are applied mechanically to a patient they do not fit, or when a minority concern is not voiced firmly enough to survive the meeting. In model panels, it arises for structurally analogous reasons: shared training lineage, anchoring on whichever position is seen first, convergence on the same

retrieved passage, and the general tendency of these systems to agree with what is already on the table.

The clinical situations where this matters are recognisable. A borderline lymph node. An indeterminate cytology. A frail patient with competing risks. A cancer that may warrant aggressive treatment or may be safely observed. A guideline that fits the average patient but not the individual in front of you.

In these situations truth is not found by averaging opinions. It is approached by testing assumptions. That is the specific capability this architecture is trying to build.

3.3 Distributed knowledge and the limits of central reasoning

The argument for plurality is not only clinical. Hayek's central observation about complex systems was that the relevant knowledge is distributed: no single institution, expert or authority holds all of it, because individuals respond to local realities, partial signals and changing conditions that cannot be fully aggregated into one centre. Attempts to centralise that knowledge can destroy the very diversity of information that makes intelligent adaptation possible.

A murmuration of starlings makes the same point in a different register. Thousands of birds move in coordinated patterns with no bird directing the flock and no central intelligence issuing commands; each responds to its immediate neighbours under a small number of simple rules. The result is not chaos but a form of collective intelligence that no single bird and no external controller could produce.

A single model that resolves disagreement into one confident answer is not a murmuration. It is one bird flying alone, telling the others where to go.

This is why a single model does not only centralise computation. It can centralise worldview. Every model reflects choices about what it was trained on, what was excluded, which values were embedded and which assumptions became invisible through repetition. Bias then appears not as obvious factual error but as omission, over-smoothing, excessive deference to consensus, and the quiet disappearance of minority interpretations.

Pluralism by design does not eliminate that. It makes it contestable, inspectable and harder to hide behind a fluent answer.

3.4 The tyranny of the majority, inside a reasoning system

If three agents examine a case and two agree, the majority is not thereby correct. In medicine the minority concern may identify the missed diagnosis, the unsafe assumption or the overlooked patient factor that matters most.

Mill and de Tocqueville both warned that majority opinion can itself become a form of tyranny when it suppresses dissent and individuality. A reasoning system that treats convergence as truth does not avoid that problem by being mechanical. It automates it — and it does so invisibly, at scale, with the appearance of objectivity that arithmetic confers.

This is the origin of the architecture’s most consequential rule: **no majority-vote mechanism, anywhere**, and no scoring scheme that reconstructs one by proxy (Section 9.5, Section 9.7).

3.5 Truth approximation through structured disagreement

In medicine, absolute truth is rarely available at the moment a decision must be made. Clinicians work with evidence, experience, imaging, pathology, patient preference, uncertainty and risk, and truth is approximated by testing these against one another.

Science progresses the same way — not through consensus alone, but through hypothesis, challenge, contradiction, revision and refinement. Shaw’s remark that progress depends on those unwilling to adapt themselves to the world is not a claim that every contrarian is right. It is a claim that progress requires space for dissent.

The design consequence is a three-way distinction that runs through the whole system:

Unstructured disagreement	is noise
Suppressed disagreement	is dangerous
Structured disagreement	can become intelligence

Truth approximation does not require universal agreement. It requires disagreement to remain visible, structured and accountable.

3.6 Reasoning sovereignty

Sovereignty in digital systems is usually discussed as infrastructure: where data is stored, who controls access, which legal framework applies, how systems are secured. Those questions matter, but sovereignty is not guaranteed by local storage. A society can host its own servers and remain intellectually dependent on opaque external models, imported institutional logic and centrally filtered representations of its own reality.

Sovereignty of reasoning therefore means retaining accountable control over how evidence is selected, challenged, weighted and presented, even where external foundation models are used. It is not a claim of freedom from foreign technology, guaranteed absence of bias, or automatic legal compliance.

The concrete requirements are: model substitutability without redesign; inspectable protocols rather than opaque prompts; documented access and egress controls; and an executable exit path from any single provider or hosting arrangement. Local ownership and governance are distinguished from physical data location; neither by itself establishes compliance.

3.7 Proportionality, reversibility, and the checkpoint before irreversible action

The purpose of clinical reasoning is to compare defensible futures — not to maximise treatment intensity, and not to pursue de-escalation regardless of risk. Meaningful outcomes include survival, recurrence, functional preservation, treatment toxicity,

lifelong medication dependence and surveillance burden. The appropriate balance is patient-sensitive.

There is a moment before a major operation or an irreversible treatment when discussion has taken place, evidence has been reviewed, and clinician and patient must commit. That moment matters, and it does not appear by accident. It is produced by the structure of reasoning that preceded it. When disagreement has been surfaced, examined and recorded — when minority concerns were voiced rather than smoothed away — the people making the final decision know what they are committing to and what remains unresolved.

Structured disagreement is not an obstacle to decision. It is what makes the decision honest.

This is a design goal, not a mandatory delay. Observation and further investigation preserve options, but delay can itself cause harm, and the most dangerous systems are not the slow ones — they are the ones that move confidently into irreversible action without having preserved space for reflection. The system must therefore display both sides and escalate where the consequence of delay is uncertain or serious (T12, Section 9.10).

3.8 Three principles, and their reach beyond medicine

Three commitments follow from the above and are traceable into specific mechanisms elsewhere in this document.

Principle	Where it becomes a mechanism
Transparency over fluency — a polished answer is not enough; a system should show what it knows, what it does not, and where uncertainty remains	Terminal states including insufficient information (§9.11); null-safe trade-offs (§9.6); decision pivots (§9.9)
Human responsibility over machine autonomy — the system supports professional judgement; responsibility stays human and inspectable	Position-before-reveal review (§9.14); clinician disposition in the audit record (§9.15); no autonomous output (§5)
Structured disagreement over false consensus — a system that hides disagreement may hide the signal that matters most	Blind first round (§8.2); consequence-graded minority protection (§9.8); sealed two-pass safety review (§9.13)

These are not cosmetic design choices. They determine what the system does when it is uncertain, which is the only moment at which a decision-support tool is genuinely tested.

The same principles apply outside medicine, and the architecture is not clinically specific. AI systems are entering public administration, cybersecurity, education and institutional decision-making, and in each the danger is not only technical failure but institutional overconfidence. A system used in public administration should show what data it relied on and where human judgement remains required. A cybersecurity system

should not suppress a minority threat signal because most indicators look normal. An educational system should not treat a child as a statistical average. Medicine is the first domain because its governance, its evidence standards and its tolerance for irreversible error make it the hardest honest test — not because the problem is confined to it.

Small sovereign jurisdictions may have a particular opportunity here: not because small countries automatically build better technology, but because governance, enterprise, clinical practice and civic responsibility can remain close enough together to test carefully and to build governance alongside innovation rather than after it.

3.9 Execution environment and infrastructure collaboration

Reasoning sovereignty has to be implemented somewhere. Phase I runs on private rented compute with a bounded set of approved model endpoints; there is no inference or gateway hardware purchase in the development baseline, because the research question is the moderation layer, not the ownership of a GPU estate.

AITEK SOVEREIGN CLOUD GROUP (ASC GROUP), SLU is the Andorran sovereign-AI infrastructure and orchestration collaborator for this work. Its supplied materials describe ASC as a sovereign and private execution environment built on EU sovereign cloud regions, with exact regions and data flows fixed contractually, and Aitek 8 as a model-agnostic orchestration and operational-intelligence layer supporting private AI, governed context, controlled model choice and agentic execution.

The shared objective is a bounded proof of concept: whether a representative Peregrine.Reason multi-agent clinical reasoning workload can execute in that environment with controlled model selection, governed context, access and egress controls, auditable execution, consumption metering and demonstrable portability. Peregrine.Reason supplies the synthetic clinical cases, clinical-agent definitions, evidence policy, Peregrine Moderator Engine, two-pass safety logic and acceptance tests.

Two boundaries are load-bearing and are stated here rather than left to be inferred:

- **Infrastructure does not acquire clinical authority.** Orchestration components may route, contextualise, meter and log, and may execute tasks within configured authority. Within Peregrine.Reason they are treated as infrastructure. They do not define clinical policy, replace Peregrine clinical agents or the moderator, or acquire clinical decision authority unless separately specified and validated. Context compression or platform substitution cannot silently alter clinical facts or dissent.
- **Portability is a requirement, not a courtesy.** The architecture must remain executable on an alternative approved configuration, and any deployment change must preserve the frozen evaluation interfaces.

The collaboration records intent and proposed contributions. It is not a completed integration, a production deployment or a performance claim, and none is asserted here.

4. The clinical problem

Guidelines, protocols and multidisciplinary team meetings improve safety, consistency and accountability. They do not remove the need to interpret evidence for an individual patient.

Difficult cases frequently involve borderline imaging, indeterminate pathology, uncertain operative risk, competing defensible pathways, incomplete data, patient-specific priorities, and thresholds for intervention that guidelines do not resolve.

Three observations frame the problem the architecture addresses.

Variation exists at team level. A UK study of therapeutic approach to low-risk differentiated thyroid cancer surveyed 75 respondents; 71 reported core MDT membership and 50 reported use of British Thyroid Association guidance, yet the extent of surgery for low-risk disease remained variable. This demonstrates variation in reported practice. It does not establish that every difference was inappropriate, nor that an AI intervention improves outcomes. The appropriate response is to examine the reasons for disagreement, not to enforce uniformity.

Variation also exists at guideline level. A published review of national thyroid-cancer surgery guidelines records significant differences between them despite common themes, attributing the persisting lack of consensus to variation in clinical practice patterns, differing appraisal of the same evidence, cultural and geographical considerations, and available diagnostic and treatment resources. Across the American, British, German and Japanese guidance compared, the recommended threshold for total rather than partial thyroidectomy and the acceptance of active surveillance differ materially. A separate 2026 AGREE II appraisal comparing the historical 2015 ATA guideline with Danish, Swedish and Norwegian guidance returned overall scores of 83%, 50%, 44% and 28% respectively, with recommendations differing especially on surgical extent and radioiodine. These are guideline-appraisal results, not probabilities that a recommendation is correct.

Reasoning is rarely recorded in an inspectable form. Some guideline variation is defensible and reflects genuine differences in epidemiology, iodine status, regulation and resources. The problem is not that one national position is correct. It is that the reasoning behind a given recommendation, and behind a given team's departure from it, is seldom recorded in a form that allows the difference to be examined at all.

4.1 The evidence ceiling

There is a further point that any system built on clinical guidance has to confront honestly, and it is easy to miss.

In the AGREE II appraisal above, the domain on which **all four** guidelines performed poorly — including the highest-scoring one — was Rigour of Development: systematic evidence searching, transparent selection criteria, explicit grading, and the link

between evidence and recommendation. That is the domain that distinguishes a guideline from an expert opinion document.

When the most heavily resourced and most widely cited guideline scores poorly there too, the implication is not merely that some guidelines could be improved. It is that **the evidence base underpinning these management recommendations is thinner than the confident language of the recommendations suggests.**

Guideline groups, however rigorously constituted, cannot manufacture evidence that does not exist. Outside a small number of areas with genuine comparative trial data, most major decisions in this disease — extent of surgery, indications for central neck dissection, TSH suppression targets, surveillance duration — rest on observational data, registry analyses and expert consensus. A guideline produced through an exemplary process may still consist largely of “consider”-level recommendations, because that is an honest reflection of what the evidence supports.

This is the design constraint, stated plainly. A reasoning system that treats guideline text as settled fact will systematically misrepresent the strength of what it is citing. It will convert “consider” into “recommended”, and weak observational support into apparent certainty, at exactly the decision points where clinicians most need to know how thin the ground is.

Three consequences run through this architecture:

- **Recommendation strength is carried as data, not lost in retrieval.** The corpus register records evidence type and recommendation strength per source (§7.2), and a matching citation is never treated as proof that the passage supports the claim (§9.4).
- **Divergence is not the failure; unexplained divergence is.** Local and national adaptation of international guidance is necessary — it is how global evidence is translated into systems that differ in resources, structures and populations. The deficiency identified in the literature is that adaptations are frequently made *without articulating why*, which forfeits the credibility that rigour is meant to provide and gives clinicians no basis for judging whether the adaptation fits their setting. The local adaptation record (§7.2) exists to make each departure, and its stated rationale, inspectable.
- **Transparency about uncertainty is itself a marker of quality** — and one that no single appraisal domain captures. A system built on this evidence base should be judged the same way.

The practical stakes are not theoretical. A UK service-level analysis reported alongside that appraisal described a department applying a nodule policy that, unlike size-stratified systems, imposed no size threshold for fine-needle aspiration: all indeterminate nodules proceeded to cytology, with a 46.9% diagnostic hemithyroidectomy rate and up to 20% of operations ultimately performed for benign disease. A 2026 joint UK consensus statement has since introduced size-stratified thresholds with explicit reference to the evidence driving the change. That is what transparent, evidence-linked adaptation looks like — a targeted, auditable update that

states its departure from prior practice, rather than a guideline rewritten from first principles.

A related omission is worth recording. None of the four guidelines appraised addressed active surveillance as an alternative to surgery for eligible patients, despite two decades of prospective data. A system that reasons only over what guidelines contain inherits their silences as well as their recommendations — which is one reason the option set is derived from the clinical ontology (§11.2) rather than from guideline text alone.

Single-model AI systems introduce a different limitation: fluent, confident output that does not expose the disagreement, uncertainty, alternative pathways or missing data underneath it.

Peregrine.Reason is designed to address that gap. It does not attempt to replace the MDT. It aims to improve the preparation, transparency and auditability of the reasoning available to clinicians before and during MDT review.

5. Intended use and boundaries

Peregrine.Reason is intended as a clinical reasoning support architecture. It is designed to structure disagreement, expose uncertainty, identify missing information, surface competing pathways and make clinical argumentation auditable.

It is **not** intended to replace clinician judgment, provide autonomous diagnosis, independently determine treatment, replace MDT governance or clinical guidelines, function as an autonomous tumour board, or operate without clinical oversight.

Explicitly out of scope for the Phase I build:

- autonomous clinical decision-making of any kind;
- direct patient-facing output;
- image interpretation — imaging findings enter as reported text only;
- prescribing, dose calculation or order generation;
- real-time or emergency use;
- live electronic health record integration;
- paediatric disease, pregnancy and urgent airway management.

Software intended for a medical purpose may fall within medical device software frameworks depending on intended purpose, functions, claims, users and target jurisdiction. Peregrine.Reason maintains deliberately narrow positioning during development and avoids claims implying autonomous diagnosis, autonomous treatment selection or approved device status. A description as “research prototype” does not itself determine device status (Section 16).

6. System overview

The reasoning flow is:

Case representation

- Independent agent assessment (blind first pass)
 - Structured challenge and revision
 - Peregrine Moderator Engine
 - Sealed two-pass independent safety review
 - Rule-based disposition and terminal state
 - Clinician decision map + audit

Component boundaries and their control constraints:

Component	Function	Control boundary
Case preparation	Extract facts, provenance, missingness and context	Clinician-defined schema; no silent filling of clinical gaps
Independent agents	Analyse the case before peer exposure	Separate execution contexts; recorded model and prompt identities
Evidence layer	Retrieve versioned, rights-cleared guidance and literature	Identical passage set to every agent; corpus version recorded per run
Deterministic anchor	Execute a bounded subset of explicit encoded rules	Rule-traceable output; ambiguous guidance stays ambiguous
Structured challenge	Surface decision-material conflicts and request revision	No instruction to follow a numerical majority; originals retained
PME	Construct claims, evidence links, trade-offs, pivots and a decision map	Explicit selection and terminal-state rules, not an opaque winning-agent score
Safety review	Challenge omissions, unsupported claims and unsafe framing	Separate execution; complete pre-filter concern register
Clinician interface	Display options, conditions, uncertainty and safety issues	No autonomous orders; no direct patient-facing release
Audit and metering	Record transformations, access, versions and costs	Tamper-evident, access-controlled, defined retention

The architecture is pluralistic by design. No single model is treated as the source of truth — including the moderator.

The design envelope is a significant freedom, and it is spent deliberately. MDT preparation is not a real-time task. The working assumption is on the order of thirty minutes per case with overnight batch processing acceptable, sized to a regional MDT workload; failure degrades to the standard MDT process rather than to a clinical

incident. These are workloads to benchmark, not a service-level commitment or achieved performance. The point of stating them is that the absence of a latency requirement is what makes multiple reasoning rounds, targeted revision and two full safety passes affordable at all. A system that had to answer in seconds could not do any of it — and would have to collapse uncertainty to appear useful. Additional safety passes and failed or retried runs are included in resource measurement rather than excluded as overhead.

7. Case representation and the evidence layer

7.1 Case representation

Clinical input enters through a versioned structured case representation defined by clinicians, not inferred by a model. Every extracted field carries a provenance pointer to its source.

A clinical schema must distinguish **absent**, **not measured**, **not applicable** and **uncertain**. A negative test is not the same as a missing test. Fields that cannot be extracted with confidence are marked missing rather than inferred, and the resulting brief lists missing fields explicitly — an incomplete workup is itself a finding an MDT wants to see.

Cross-object validation prevents category errors that are easy to make and consequential: a postoperative thyroglobulin observation being used as a preoperative fact; a recurrence-risk estimate being labelled as mortality risk; model confidence being recorded as a calibrated probability; a prognostic estimate being read as a causal treatment effect.

7.2 Evidence corpus

A versioned, controlled corpus registers each source with its version, jurisdiction, population, evidence type, recommendation strength, licence status and superseded status. The 2025 ATA adult differentiated thyroid cancer guideline is the contemporary thyroid backbone for corpus planning, with applicable corrigenda tracked and the historical 2015 guideline separately labelled where used to explain earlier practice. Other guidance is included subject to rights clearance; no proprietary source is a completion dependency.

Retrieval supplies the **same passage set to every agent**, so that citation availability is not itself a source of inter-model variance. A corpus version bump is a re-validation trigger, not a routine update.

A **local adaptation record** links a source recommendation and version to any changed local recommendation, its supporting evidence, the resource or patient-context rationale, the responsible approver and a review date. Neither a reference count nor an overall guideline-appraisal score substitutes for assessing the particular recommendation and its applicability to the case in hand.

7.3 The deterministic anchor

Alongside the model panel, a **deterministic guideline and rules anchor** encodes a bounded subset of explicit recommendations, with versioned exceptions and explicitly indeterminate branches.

Its value is that it cannot hallucinate. Divergence between the anchor and the model panel is a high-value review signal in both directions: either the panel has identified a case-specific reason to depart from the encoded rule, or it has drifted.

The anchor is not treated as ground truth. Source interpretation, encoding defects, outdated rules and inapplicable inputs can all cause it to fail. Ambiguous guidance must remain ambiguous rather than being silently forced into an exact rule, and anchor–model divergence is never automatic proof that the model is wrong.

8. Independent agent layer

8.1 Panel composition

The Phase I baseline is **three distinct model families plus the deterministic anchor**. This is a deliberate feasibility choice, not a statement that three is optimal. Expansion to a larger panel is permitted only where measured marginal value, capacity and cost justify it.

Panel diversity is the mechanism by which the architecture is supposed to generate value. Two models sharing a training lineage and an alignment methodology produce correlated errors, and correlated errors are invisible to any aggregation scheme. The panel is therefore constructed across four axes: training corpus, national and regulatory origin, alignment methodology, and open versus closed weights.

Candidate deployment modes include locally executed open-weight models with no network egress and externally hosted frontier endpoints under contractual data controls. Specific checkpoints are not fixed in this document: each deployed checkpoint requires its own licence, provenance, security and performance record, and local execution without egress differs materially from sending cases to a foreign-hosted service without removing supply-chain or institutional-policy risk.

Two operating requirements follow:

- **Version pinning is mandatory.** Closed API models change without notice. Every panel member is pinned to a specific version string, and a version change is a re-validation trigger.
- **The panel must degrade gracefully.** Loss of any arm — for legal, contractual or institutional-policy reasons — must be a configuration change, not a redesign. Results are reportable with and without each material arm.

8.2 Interrogation protocol

Round 1 — blind independent assessment. Every panel member receives the identical case representation and answers the full question set without sight of any other member’s output. This is non-negotiable: allowing models to see each other’s answers before forming their own produces anchoring cascades and manufactures false consensus.

Round 2 — role-scoped assessment. Members are additionally queried under specialty frames — thyroid surgery, endocrinology, imaging, pathology and cytopathology, nuclear medicine, evidence appraisal, and patient values — to elicit considerations a general prompt suppresses. Advanced oncology, genetics and perioperative-risk frames are activated only where the case requires them.

This round carries a known risk and is treated as a hypothesis: persona prompting can degrade factual accuracy. Round 2 is ablated during evaluation and dropped if it does not demonstrably add value over Round 1.

Round 3 — bounded adversarial pass. Selected members are shown the anonymised set of Round 1 positions and asked solely to identify factual errors and unstated assumptions. They are explicitly instructed **not** to revise toward the majority. Round 3 output enters PME as critique, never as a replacement position.

Each member returns a structured opinion containing a primary position mapped to a decision node, ranked alternatives, explicit reasoning, evidence citations, stated uncertainty, safety flags, and a list of information that would change the answer. Free-text-only reasoning is insufficient: what cannot be structured cannot be scored.

8.3 Independence of execution is not independence of errors

Separate execution contexts do not produce statistically independent errors. The programme therefore **measures** correlation rather than assuming it away: pairwise decision agreement, joint error rates, conditional co-error rates, and leave-one-out change in performance. Cases on which all models agree *incorrectly* are explicitly retained for analysis, because they are the failure mode that aggregation cannot detect.

The common evidence bundle controls for citation availability but can itself induce correlated error; a bounded retrieval-ablation experiment addresses this after the primary comparison.

No universal correlation threshold is asserted as establishing independence. The useful question is narrower and answerable: does the additional arm improve the pre-specified outcome without unacceptable burden?

9. The Peregrine Moderator Engine

9.1 Why the moderator is the decisive component

Without a strong moderation layer, a multi-agent clinical reasoning system becomes one of four things: a noisy collection of competing outputs; a disguised single-model answer that smooths disagreement away; an unsafe consensus machine; or an apparently sophisticated system that still hides uncertainty from the clinician.

A weak moderator makes a multi-agent system unsafe even where the individual agents are strong. Moderator behaviour is therefore treated as a clinical-safety-critical component and evaluated separately from the agents.

9.2 What PME is not

PME does **not**:

- compute a mean opinion, a majority vote or a smoothed synthesis;
- choose a winning agent;
- assign global authority to an agent because it performed well elsewhere in the case;
- force agreement, or suppress dissent for neatness;
- derive a preference by summing ordinal scores into a master score;
- act as a central source of clinical truth;
- introduce unlabelled new clinical claims;
- make autonomous clinical decisions or generate treatment instructions.

The purpose of the architecture is to reduce dependence on any single model, institution or reasoning pathway. A moderator that replaced plural agents with one dominant central judgment would recreate exactly the failure mode the system exists to avoid.

PME must therefore itself be governed: transparent rubrics, version-controlled protocols, auditable synthesis logic, separate safety moderation, human review, reproducible argument graphs, clear separation between agent reasoning, moderation, safety review and clinician decision, and the ability to compare different moderator implementations during evaluation.

Moderator self-preference is a measured risk. Judge models systematically favour outputs resembling their own. The moderator is therefore either drawn from a family not otherwise represented in the panel, or run in a configuration where panel outputs are stripped of stylistic signature before synthesis. Self-preference rate is measured directly during evaluation by injecting opinions of known provenance.

9.3 Execution pipeline

PME is specified as a sixteen-stage pipeline with an explicit implementation allocation per stage. Deterministic components enforce recorded rules; model-assisted components propose interpretations that remain traceable and challengeable. A

deterministic gate can still act on an incorrect model label, so extraction, classification and rule execution are tested separately.

#	Stage	Allocation and required behaviour
0	Session integrity	D — validate schemas, manifest, access and session state; fail closed on invalid prerequisites
1	Atomic claim extraction	H — model-assisted extraction; deterministic IDs, source spans and schema checks
2	Case and evidence verification	H — deterministic source/locator checks; model-assisted entailment with unresolved status preserved
3	Argument graph	H — model-assisted edge proposals; deterministic typed graph and reference integrity
4	Disagreement classification	M — classify with evidence-linked rationale; no authority from the label alone
5	Clinical trade-offs	H — model-assisted evidence extraction; deterministic units, horizon checks and null-safe arithmetic
6	Epistemic review	M — evidence, applicability and uncertainty assessments with provenance; no scalar winner
7	Clinical consequence review	M — materiality and consequence against the approved outcome register
8	Minority protection	H — model-assisted plausibility and severity labels; deterministic escalation rules and audit
9	Decision pivots	H — model-assisted identification; deterministic links to facts, conditions and options
10	Targeted challenge and revision	H — rule-based routing and iteration bounds; responses stored without overwriting originals
11	Conformity and duplicate checks	H — exact duplicate and integrity checks plus tested semantic and conformity detectors
12	Constrained synthesis	H — rendering constrained to accepted objects; deterministic content and trace checks
13	Separate safety review	H — blind first pass on the full claim register; reconciliation pass with reasons; both retained
14	Rule-based disposition	D — execute approved admissibility, dominance, terminal-state and release rules
15	Clinician output and audit	H — rule-bound display, versioned audit and metering; disposition remains human

D = deterministic control · **M** = model-assisted interpretation · **H** = hybrid.

Clinical and methodological authorities approve the schemas, thresholds and reference labels. No model may silently redefine them.

9.4 Core object contracts

Moderation operates on typed objects, not on narrative text. The unit of moderation is the claim or reasoning contribution, not the agent — an agent may make one strong and one weak contribution in the same case.

Object	Essential fields and invariant
ClaimObject	Source agent, round and span; claim type; decision node; case facts; assumptions; evidence IDs; related claims; uncertainty. <i>Every material claim retains its original provenance.</i>
EvidenceObject	Source and version; locator; population; outcome; horizon; study type; quality and applicability; rights. <i>A matching citation is not proof that its passage supports the claim.</i>
DisagreementObject	Constituent claims; disagreement class; clinical importance; missing information; resolvability. <i>Synonyms alone do not create a clinical conflict.</i>
ClinicalTradeoffObject	Option and comparator; separate outcome estimates; baseline and treated risks; horizon; harms; uncertainty; reversibility; patient and provider modifiers. <i>Null is a valid value.</i>
MinoritySignalObject	Claim IDs; plausibility; severity; reversibility; resolvability; escalation level and action. <i>Supporting-agent count is descriptive only.</i>
DecisionPivotObject	Variable; present state; alternative state or threshold; affected pathways; evidence; required data. <i>A hypothetical pivot is not an observed fact.</i>
RevisionObject	Original and revised position; the challenge; evidence or rationale for change; persistent dissent. <i>Originals are never overwritten.</i>
SafetyReviewObject	Blind first-pass findings; reconciliation findings; hidden and revealed dispositions; missed concerns; changes and reasons; final disposition. <i>No overwriting and no automatic PASS.</i>
ModeratorOutput	Decision state; admissible pathway set; policy and rule trace; pivots; safety disposition; evidence links; interface visibility settings. <i>Counts remain secondary audit information.</i>

All IDs, units, required fields, permitted nulls and enumerated values are validated by executable schemas. The serialisation format is an engineering choice; the field semantics are under joint clinical and technical change control.

9.5 Scoring: interpretable vectors, no master score

PME scores reasoning contributions across separate epistemic and clinical-consequence dimensions — evidence grounding, applicability, case consistency, uncertainty handling, missing-data detection, consequence if wrong, reversibility, patient-value dependence and safety framing.

These produce **interpretable score vectors**. They do not produce a single aggregate quality number, and they never select a pathway by arithmetic.

This is a deliberate departure from the intuitive design. Collapsing multiple ordinal dimensions into a weighted total looks transparent and is not: it hides which dimension drove the result, invites tuning until the desired answer appears, and quietly reconstructs a majority-vote mechanism when the dimensions correlate with agent count. Score vectors are shown alongside the claim so a clinician can see *why* a contribution is weak — without the arithmetic ever making the decision.

9.6 Clinical trade-off objects and benefit–harm discipline

Each management option is compared against a specified alternative, for a specified outcome, population and time horizon.

Absolute risk reduction is the difference between comparator event risk and intervention event risk. Number needed to treat is the reciprocal of a positive ARR expressed as a probability; number needed to harm uses a positive excess harm risk. These are **outcome-specific summaries, not fixed properties of an operation or a drug**.

The discipline that makes them safe to compute:

- A small denominator for a temporary adverse effect cannot be compared mechanically against a large denominator for prevention of death. Severity, duration and patient values matter, and NNT and NNH are never divided into an automatic recommendation score.
- Where effects are unsupported, incompatible or cross the no-effect boundary, the numerical result is withheld or explicitly qualified. An apparently precise number must not conceal unresolved uncertainty.
- Where quantitative prerequisites are absent or incompatible, ARR, NNT and NNH remain **unestimated**, with the reason stated. The system does not invent a number because a model remembers similar ones.
- Treatment burden is itself an outcome. Total thyroidectomy produces complete reliance on exogenous thyroid hormone; lobectomy may retain endogenous function, with or without supplementation. Dependency, monitoring, symptoms and patient-reported outcomes are recorded without assuming that every patient has worse quality of life after total thyroidectomy.

How this is tested, concretely. A software arithmetic fixture supplies hypothetical risks of 4% without an intervention and 3% with it, for the same outcome over a defined ten-year horizon: ARR is 1 percentage point, so NNT is 100. A separate hypothetical

permanent-harm excess of 2 percentage points gives NNH 50. These are invented test numbers, not clinical estimates.

The test passes only if the output preserves the comparator, endpoint, horizon and hypothetical status, computes the arithmetic correctly, **and refuses to conclude that treatment is favourable or unfavourable by comparing 100 with 50**. That refusal is the actual requirement. Two numbers that are dimensionally similar and clinically incommensurable must not be ranked against each other, and a system that does it once will do it silently thereafter. Further fixtures cover proportions versus percentages, zero and negative effects, incompatible horizons, and uncertainty intervals spanning no effect. In a live case with inadequate effect evidence, the values return unestimated even where a model recalls similar figures from training.

Material outcome families, each compared with its own discipline:

Outcome family	Required comparison discipline
Disease-specific survival	Mortality over a stated horizon; never inferred from recurrence or biomarker improvement
Clinically meaningful disease control	Structural recurrence or persistence, local progression, salvage need — each separately defined
Serious or permanent treatment harm	Permanent nerve or parathyroid injury, major toxicity and other irreversible decision-relevant harms
Temporary harm and treatment experience	Duration, recovery, symptoms, procedure burden — not merged with permanent disability
Function and quality of life	Voice, swallowing, airway, endocrine function and dependence, validated patient-reported outcomes where available
Competing health risk and burden	Comorbidity, competing mortality, monitoring, repeated treatment, and the consequences of delay
Patient values and reversibility	Recorded priorities, acceptable trade-offs, follow-up feasibility, availability and cost of later escalation

Patient values are **recorded, not inferred** from demographics and not encoded as universal utility weights. All potentially decision-changing safety outcomes remain material, and the synthesiser cannot drop an inconvenient dimension in order to manufacture a clean comparison.

9.7 Admissibility, and how a pathway can be presented as preferred

PME first determines admissibility, then whether a preference is supportable at all.

First, each option is classified as *supported*, *conditional*, *unsupported* or *unassessable* against documented case facts, applicable evidence, unresolved safety concerns and material missing information. “Unsupported” does not mean false: an unverified safety concern remains visible for review even where it cannot justify a positive treatment preference.

Second, no pathway may be preferred while an unresolved high-consequence safety objection or a decision-critical missing fact could reverse that preference. If exactly one option remains supported after this check, PME may report it as the only currently supported pathway and show why the alternatives failed the stated tests.

Third, where more than one option remains defensible, a provisional preference is permitted **only under a documented dominance test**: the option is no worse on every pre-specified material outcome and meaningfully better on at least one, using comparable outcomes, horizons, evidence and clinically ratified tolerance rules.

For a quantitative outcome oriented so that higher is better, let $\Delta = \text{outcome}(A) - \text{outcome}(B)$. The approved policy supplies a no-worse tolerance ε and a minimum meaningful improvement δ . Dominance requires the uncertainty-aware lower bound of Δ to be at least $-\varepsilon$ for every material comparable outcome, and greater than δ for at least one. Categorical comparisons require explicit, evidence-supported order or equivalence rules. **Where an approved tolerance is absent, the superiority condition is not met** — the programme does not invent universal clinical thresholds.

Null-safe by construction. A null or non-comparable material estimate cannot be counted as zero harm, no difference, equivalence, or a favourable score. It blocks dominance unless the policy documents a justified qualitative comparison or explicit non-applicability. This does not by itself establish equipoise: two or more supported pathways must still remain. Higher-severity safety gates take precedence over any numerical dominance result.

The rule is deliberately conservative, and the asymmetry is intentional: **inability to demonstrate dominance is not evidence that an option is inferior**. Where several adequately supported pathways remain, PME preserves them. Missing treatment-effect evidence may prevent ranking without preventing a useful discussion; missing decision-critical case information instead triggers the relevant missing-information state.

These definitions live in a versioned **Clinical Decision Policy Record**, owned by the clinical lead, independently reviewed, and frozen before any scored evaluation run. It is not replaced by a model's unrecorded preference.

9.8 Minority-signal protection

In clinical reasoning, the minority concern may be the signal. A single member correctly flagging medullary carcinoma on an unexplained calcitonin is among the highest-value outputs the system can produce — and simple consensus logic would delete it.

Minority signals are graded by **consequence if missed**, not by how many agents raised them:

Level	Meaning	Action
0	Immaterial difference, duplicate wording or unsupported noise	Retain in audit; do not generate a clinical alert merely for diversity
1	Plausible low-consequence alternative	Preserve where relevant; no mandatory urgent escalation

Level	Meaning	Action
2	Plausible management-changing alternative	Show the alternative and its supporting conditions
3	Plausible serious safety concern	Targeted challenge and safety review; preserve until explicitly resolved
4	Plausible catastrophic or irreversible harm if missed	Human escalation unless satisfactorily resolved and documented

Severity alone cannot promote arbitrary speculation into a diagnosis: plausibility, case support and uncertainty are assessed explicitly. Conversely, a concern need not carry a guideline quotation to deserve investigation when a case fact suggests serious risk.

9.9 Decision pivots

A decision pivot records the specific variable whose change would move the decision: its present state, the alternative state or threshold that would matter, the pathways affected, the supporting evidence, and the data required to resolve it.

This is what converts “we are uncertain” into something a clinician can act on. “Uncertain nodal status” is an observation. “If the borderline level VI node is confirmed pathological, the central-neck management option set changes, and the resolving investigation is X” is a decision.

A hypothetical pivot is never recorded as an observed fact.

9.10 Hard trigger rules

The following controls prevent unverified material from being treated as clinical truth. They execute deterministically.

ID	Trigger	Required action
T1	Material claim lacks verified support	Mark unverified; exclude as a positive basis for preference; retain safety implications and audit
T2	Citation fabricated or unverifiable	Remove that citation’s authority; log as a model-level defect; seek independent support where appropriate
T3	Missing information could reverse a serious decision	Block unconditional preference; report insufficient information or escalate
T4	Level 3 or 4 concern raised	Targeted challenge and separate safety review; no majority suppression
T5	Synthesis introduces a new claim	Label as moderator inference; validate separately or exclude from the released rationale
T6	Quantitative prerequisites absent or incompatible	Leave ARR/NNT/NNH unestimated; state why; do not invent a number
T7	Material conflict between applicable guidelines	Show source provenance, strength and applicability; no citation-count tie-break

ID	Trigger	Required action
T8	Unsupported capitulation, or un rebutted dissent disappears	Preserve the original; flag conformity; remoderate where material
T9	Safety reviewer requests remoderation	Withhold preferential output; perform a bounded corrective pass
T10	Unresolved sentinel concern or human-escalation state	No automatic preference; display the concern and the clinician-resolution requirement
T11	Critical integrity or provenance failure	Stop ordinary moderation; distinguish technical failure from clinical uncertainty
T12	Irreversible option with a defensible less-irreversible alternative	Explicitly compare delay, salvage and patient preferences before any preference
T13	More than one supported option, no dominance rule satisfied	No automatic preference. Null material estimates cannot prove equivalence. Preserve supported alternatives or return conditional/insufficient information
T14	Invalid object, failed required safety review, or absent audit prerequisites	Fail closed; do not label the episode a successful clinical output

Technical failure is scored as technical failure. Sessions carry a separate status — valid, degraded or failed — and a technical fault must never be presented as a clinically appropriate finding of insufficient information. That substitution would make the system’s apparent clinical calibration improve every time its engineering got worse. Repeated remoderation failure ends in a logged technical or safety escalation, not an unbounded loop.

9.11 Terminal states

Every episode terminates in exactly one of five states, in this precedence order:

1. **Safety escalation** — an unresolved serious danger requires human resolution.
2. **Insufficient information** — a safe clinical distinction cannot yet be made.
3. **Conditional convergence** — a specified unresolved condition is decisive; the map shows what resolves it.
4. **Structured equipoise** — multiple options remain supported and no dominance rule is satisfied.
5. **Structured convergence** — the explicit single-option or dominance rule is satisfied.

Insufficient information and structured equipoise are correct outputs, not failures.

In medicine, false certainty is more dangerous than acknowledged uncertainty.

Genuine evidence limits are preserved even when every agent agrees.

9.12 Anti-convergence and anti-gaming controls

The system is tested against the ways it could produce the appearance of good reasoning:

Control case	Correct system behaviour
Different words for lobectomy / hemithyroidectomy	Normalise vocabulary; no spurious material disagreement
Duplicate an existing agent response ten times	No change in clinical authority or pathway selection from repetition
Reorder or relabel model outputs	Investigate any change in judgment; measure position and provenance bias
Fully supported ordinary agreement	Convergence permitted; no invented controversy
Unsupported catastrophic speculation with no case support	Do not escalate merely because language is dramatic
Well-supported minority alert removed by primary synthesis	Safety reviewer detects the omission from the full pre-filter register
Legitimate revision after new evidence	No conformity penalty for changing position
Small NNH for transient harm versus large NNT for preventing death	No automatic arithmetic ranking across non-equivalent outcomes
One material outcome null, another favouring an option	Do not infer equivalence from null; preserve alternatives or name the missing premise
Primary moderator supplies a persuasive but incorrect exclusion reason	Blind pass records the concern independently; reconciliation does not silently overwrite it
Same content displayed with high versus low agreement counts	Audit the interface condition and measure clinician disposition shift

9.13 Sealed two-pass safety review

Safety review is separated from synthesis, and structured so that it cannot be anchored by the synthesis it is checking.

Pass A — blind. The safety reviewer receives the case, the admissible evidence, and the **complete pre-filter claim and concern register** — including unassessed and rejected items — *without* primary retain/reject labels, exclusion reasons, support counts, model identities or primary scores. It may also identify omissions that no agent raised. Its findings are **sealed** before it sees anything else.

Pass B — reconciliation. The primary synthesis, dispositions and exclusion reasons are revealed. Changes in safety disposition and the evidence for them are recorded. Neither pass overwrites the other.

The reviewer checks for unsupported facts, missed serious concerns, mishandled uncertainty, recurrence/mortality conflation, incomparable effects, omitted harms, comorbidity, patient values and unsafe framing. It runs in a separate execution context with its own prompt and rule version and its own disposition.

Full Pass A coverage cannot be restricted to the concerns the primary moderator happened to select. **Incomplete coverage or a failed pass blocks release.** Using different model families for primary and safety roles is a preferred configuration, not a guarantee of independent errors — and it is evaluated as such, using seeded omissions and false alarms that include concerns discarded upstream.

9.14 Clinician interface and automation bias

The primary decision map displays pathways, evidence, trade-offs, pivots and safety concerns **without numerical support counts or majority badges**. Counts and full provenance are available in a clearly labelled expandable audit view; access is logged, and the interface states explicitly that repetition is not evidence.

Human oversight is architectural rather than advisory. In formative review, clinicians record their own provisional position **before** the brief is revealed, and the divergence between the two is captured as study data — which is also how automation-bias magnitude is measured rather than asserted.

The display effect of support counts is itself tested: a bounded, counterbalanced count-visible versus count-hidden comparison holds the substantive argument content constant and measures disposition change, justification quality, comprehension and review time.

9.15 Audit, reproducibility and change control

Every run writes an append-only audit record: original agent outputs, all transformations, evidence corpus versions, proposed and rejected claims, the blinded safety input and both safety passes, rule traces, display configuration, detail expansion, the rendered map, clinician disposition and divergence from the system's position. Model, node, prompt, rubric and schema versions are recorded in a pinned manifest.

Reproducibility is defined honestly. It means deterministic rule execution and measured output stability under a pinned manifest — not an unqualified promise of bit-identical inference across all arms and providers. Additional safety passes and failed or retried runs are included in resource measurement.

Because the moderator shapes how uncertainty reaches clinicians, changes to prompts, rubrics, schemas, safety filters, orchestration, graph logic, dominance rules, output formatting or model versions are version-controlled, documented and tested before deployment. The record states what changed, why, who approved it, which evaluation cases were used, whether synthesis behaviour altered and whether any

safety risk was identified. A case processed under one moderator version is not equivalent to a case processed under another until that change has been assessed.

An audit hash alone is not a reversible record; retention of reconstructible content is separately authorised and protected.

9.16 Moderator failure modes

The moderator is tested against its predictable failures: premature consensus; smoothing away disagreement; favouring fluent reasoning over correct reasoning; over-weighting majority position; ignoring minority risk signals; manufacturing disagreement to appear useful; hallucinating supporting evidence; producing false certainty; applying guidelines too rigidly; ignoring patient-specific factors; failing to detect missing data; generating hidden new claims; presenting support as instruction; becoming a centralised source of truth; producing scores clinicians cannot interpret; and failing to escalate.

If the moderator fails, the system fails. That is why moderator evaluation is a distinct workstream rather than a component of overall accuracy.

10. Why this architecture might be wrong

A whitepaper that cannot state its own failure conditions is marketing. The following are the live risks to the central thesis, each with the test that would expose it.

A single strong model may match the panel. This is the most consequential possibility. Leave-one-out ablation and a **required strong single-model benchmark** are built into the evaluation specifically to test it. A simpler system is allowed to win. If one model matches the panel, the panel is expensive theatre and the finding should be published as such.

Generic summarisation may match structured moderation. The primary comparison holds the panel outputs frozen and varies only the moderation layer, so the comparison isolates PME rather than the panel.

Dissent discovery may manufacture noise. Measured as dissent-flag precision and recall against independently adjudicated equipoise, and through the anti-gaming library in Section 9.12.

Claim extraction may not be reliable enough to support the rest. Everything downstream depends on it. Claim-extraction fidelity is a **mandatory acceptance measure**, not a secondary metric, with independently labelled reference claim units, a pre-specified matching rule and denominator, and prospective targets of at least 95% precision and 95% recall for atomic claims, with no unflagged management-changing negation, attribution, numeric, unit or safety-omission error in a released audited output. Raw and released-output errors are reported separately. An insufficient denominator is not a pass.

Panel members may fail together. Addressed by measuring correlation directly (Section 8.3) rather than assuming diversity produces independence.

The safety reviewer may share the primary moderator’s blind spots. Addressed by blind Pass A, seeded-omission testing, and different-family configuration — which is tested, not assumed sufficient.

Uncertainty preservation may make the system uselessly indecisive. Over-abstention is a failure mode in its own right. False-escalation rate is reported alongside detection rate; a system that returns “insufficient information” on everything has not solved the problem.

The added complexity may not earn its cost. Per-arm marginal value, resource consumption and cost per episode are measured.

10.1 Principal risk register

Risk	Detection or decision point	Response
PME does not outperform a strong single-model baseline	Primary paired comparison and cost report	Report the negative result; prefer the simpler architecture where justified
Hidden judge-model selection — a scalar tie-break reappears by proxy	Rule trace; duplicated and reordered opinion tests	No unversioned scalar tie-break; preserve equipoise
Minority suppression, or the opposite — excessive alerting	Full pre-filter safety review; retention and precision measures	Paired safety and false-escalation targets, reviewed clinically
Misleading benefit–harm numbers	Provenance, comparator, outcome, horizon and unit checks	Null-safe arithmetic; no NNT/NNH ranking across unlike outcomes
Correlated error across agents <i>and</i> the safety reviewer	Joint and conditional error analysis; arm ablations	Disclose residual correlation; never assume independence
Incomplete data or unauthorised transfer	Case-integrity, rights and egress tests	Synthetic-first baseline; approved processor route before any clinical data
Limited medical-software engineering experience in the team	Quality-system and reference review	Appoint a qualified adviser and independent evaluator; no unsupported certification claim
Compressed schedule or	Resourcing gate and periodic delivery review	Reallocate or defer optional scope — safety gates are never removed to protect a schedule

Risk	Detection or decision point	Response
unavailable specialists		
Infrastructure or platform dependency proves unsuitable	Proof-of-concept tests routing, audit, metering, residency and portability before broader commitment	Keep clinical logic platform-independent; retain export evidence and a qualified alternative compute route; no proprietary component becomes a clinical dependency
Fabricated citations presented as authoritative	Exact-match corpus verification (T2)	Citation authority removed; logged as a model-level defect
Silent capability drift in externally hosted models	Version pinning; continuous canary case set	A version change is a re-validation trigger, not a routine update
Overfitting to one centre's practice	Comparison across sources of reasoning	Multi-centre validation before any indication expansion
Liability ambiguity	Named clinician accountability at sign-off	System positioned as non-autonomous throughout; clinician disposition recorded

11. Phase I domain: adult differentiated thyroid cancer

The initial domain is a methodological choice, not a market boundary.

Adult differentiated thyroid cancer offers concentrated examples of surveillance, diagnostic uncertainty, treatment extent, postoperative reassessment and preference-sensitive care. Many patients have a favourable course while a smaller group requires prompt recognition of aggressive disease — so **both inappropriate treatment and inappropriate restraint can cause harm**. The dominant clinical problem is overtreatment rather than undertreatment, which makes decision-support errors less likely to be acutely catastrophic than in, say, sepsis or acute stroke pathways. Guidance is densely codified and largely textual, and decisions recur over long follow-up, so a single case generates multiple evaluable decision points.

Safety boundaries. Medullary, anaplastic and poorly differentiated disease must be recognised as important safety and specialist-routing categories. Their recognition does not authorise the differentiated-thyroid-cancer module to provide a complete management pathway outside its validated scope.

11.1 The clinical asymmetry this domain exposes

Differentiated thyroid cancer has an unusual structure that makes it an exacting test of reasoning rather than a soft one.

Long-term disease-specific survival exceeds 95%, yet rates of persistent or recurrent disease after treatment are substantially higher. The limited effect of that recurrence on survival implies something that is uncomfortable to state and important to model: successful treatment may not always require eradicating all disease. At the same time, a minority of patients have genuinely aggressive disease that is difficult to distinguish from the innocuous majority during workup.

Those two facts interact badly. The difficulty of identifying the aggressive minority creates pressure to treat the innocuous majority more intensively than they need — and because treatment morbidity scales with treatment intensity, in an anatomically confined region containing vital and functionally important structures, that pressure has a measurable cost in permanent nerve and parathyroid injury, lifelong hormone dependence and quality of life. Rising incidence attributed largely to increased diagnostic sensitivity, against stable mortality, means the absolute number of patients exposed to that trade-off continues to grow even as the proportion falls.

This is precisely a reasoning problem rather than an information-retrieval problem.

The clinician is not short of guidance. They are being asked to weigh a low-probability, high-consequence possibility against a high-probability, moderate-consequence harm, in an individual whose profile does not match the average of the cohorts the evidence came from — study populations carry selection effects, and the average profile of a published cohort frequently does not describe the patient in front of you.

Two design commitments follow directly:

- **Quantification and valuation are separate steps.** The clinician’s task is to quantify and compare benefit and harm; attributing value to that comparison is the patient’s. The system therefore produces trade-off objects with outcomes, horizons and uncertainty made explicit, and records patient priorities where they exist rather than inferring them (§9.6). Presenting benefit and harm in comparable, understandable terms is useful to practitioners and patients alike; collapsing them into a single recommendation score removes exactly the step that belongs to the patient.
- **Applicability is checked, not assumed.** Whether an estimate derived from a study population applies to this patient is a question the system must be able to raise and, where it cannot be answered, to leave unresolved (T1, T6, §9.7).

11.2 Clinical distinctions preserved in the ontology

The ontology exists so that disagreement is measurable. Free-text-only output cannot be scored, and a decision structure that collapses genuine clinical distinctions will measure the wrong thing.

Decision area	Required distinction
Observation and intervention	Whether to intervene precedes the choice of operation. Active surveillance is an organised follow-up strategy, not absence of care. Growth is interpreted with location, kinetics, other risk features and applicable guidance

Decision area	Required distinction
Nodal disease	Central versus lateral; clinically or radiologically apparent versus incidental microscopic disease; number, deposit size, extranodal extension and compartment. Microscopic positivity is not automatically equivalent to bulky disease
Neck surgery	Prophylactic central dissection, therapeutic central treatment and therapeutic lateral treatment are separate reasoning questions. The indication must not be inferred from the phrase “node-positive” alone
Extrathyroidal extension	None, microscopic or minimal, gross strap-muscle involvement, gross critical-structure invasion, or indeterminate — with the implicated structure specified
Histology and risk	Aggressive or poorly differentiated features may change the safety pathway. Disease-specific mortality staging and recurrence risk remain separate
Postoperative management	Response after surgery updates later decisions. A risk category is not an immutable instruction for radioiodine, TSH suppression or surveillance intensity
Quality of life	Voice, swallowing, endocrine function, permanent hypoparathyroidism, toxicity, medication dependence, anxiety and surveillance burden remain distinct from recurrence

Controlled decision nodes span diagnostic information, disease characterisation, observation versus intervention, surgical and nodal management, postoperative and longitudinal management, special circumstances, and cross-cutting modifiers including comorbidity and competing risk, treatment burden, decision reversibility, patient preferences and provider context.

11.3 A worked example of the discipline

Consider a postoperative radioiodine decision in which initial pathology and reassuring follow-up information appear to support different intensities of care.

The system does not begin with “should this patient have radioiodine?” It begins with: *which outcome would radioiodine improve — structural recurrence, biochemical persistence, mortality, or another defined outcome?* It then records the facts required to answer that (surgical extent and completeness, histology, nodal burden, gross extension, imaging, thyroglobulin and antibody results, and the assay, timing, TSH and stimulation context), and the evidence prerequisite (a comparative source with matching population, comparator, endpoint and horizon).

Where no such source is supplied, baseline risk, treated risk, ARR, NNT and NNH all remain **null**. The pivots are named — a material residual lesion, aggressive histology, unreliable biochemical interpretation, or a different patient preference. The rule trace records T3 for critical missingness, T6 for unsupported numbers, T7 where applicable guidance conflicts, T13 where no defensible preference rule is met. The terminal result

is insufficient information, a conditional map, or equipoise — **not an automatic radioiodine or no-radioiodine instruction.**

A negative stimulated thyroglobulin is not permitted to erase other high-risk evidence, and is not permitted to create a comparative treatment effect that was never measured. Interpretation requires surgical extent, timing, assay characteristics and quantification limit, antibody status, TSH context, stimulation method and relevant imaging; a negative result is not a universal rule to omit radioiodine, and stimulated testing is not made mandatory for every patient.

12. Evaluation

The system is evaluated on reasoning quality and safety behaviour, not on speed and not on agreement with a historical decision.

12.1 The central research question

On an independently reviewed, deliberately enriched stress-test set, does structured claim-level moderation improve five-state decision-state classification compared with generic summarisation of the same frozen panel contributions, while preserving pathway coverage, traceability and safety?

The primary endpoint is **balanced accuracy of the five-state classification** against independently adjudicated reference states, with PME compared against a generic summariser receiving identical frozen panel outputs.

This endpoint is formative and conditional on a designed case mix. **It is not a prevalence-weighted estimate of real-world clinical accuracy**, and it is not a clinical effectiveness measure.

12.2 Hypotheses and their tests

Hypothesis	Measure and comparison	Interpretation
PME improves decision-state classification	Balanced accuracy versus generic summary on identical cases and frozen panel outputs	Primary formative comparison against independently adjudicated reference states
Material dissent is retained without being manufactured	Retention, precision, false-consensus and false-escalation rates	Sensitivity and over-alerting assessed together
The panel adds useful information	Strong single-model baseline; pairwise error overlap; leave-one-out	More arms are justified only where measured value supports them
Trade-off and pivot structures improve comprehension	Option and pivot coverage; blinded clinician comprehension tasks	Evaluates usefulness, not clinical effectiveness

Hypothesis	Measure and comparison	Interpretation
Separate safety review detects additional failures	PME with and without safety review on stored outputs	Incremental detection <i>and</i> additional false alarms both reported

12.3 Design controls

- **Independent case authorship and adjudication.** Cases are authored, labelled and adjudicated by people independent of the build team, with reference labels and analysis plan locked before scored outputs are opened. No reviewer labels their own authored case.
- **A sealed hidden challenge partition** that can block progression after the primary report is frozen.
- **A pre-specified minimum case count per decision class.** If the minimum is not met, the named secondary is descriptive case-level exact agreement — reported as such, never relabelled as a successful primary.
- **Safety is non-compensatory.** Mandatory safety and provenance failures cannot be offset by a good aggregate score, and no global quality score is permitted to hide the component trade-offs.
- **The analysis plan is frozen before unblinding.** A margin for material research disadvantage is pre-specified and applies to PME versus generic summarisation *and* to PME versus the strong single model. It is not a clinical non-inferiority margin.

12.4 Progression decision

Phase I terminates in a signed progression record, not a marketing claim.

GREEN — all mandatory safety, provenance and traceability gates pass; no unresolved sentinel event; class and denominator minimums met; PME not materially disadvantaged against either required comparator under the frozen margin; and the totality of pre-specified evidence, including interval width, supports a larger independent study. GREEN authorises designing and costing that study. It is not a claim of clinical superiority, non-inferiority or permission for patient care.

AMBER — safety gates pass but precision, class minimums, mapping reliability, benchmark or usefulness evidence do not justify unchanged progression. The locked result is retained; targeted refinement or a separately funded precision study follows. An unassessable primary is not described as successful.

RED — an unresolved sentinel failure, a mandatory safety or provenance failure, or convincing material disadvantage against a required comparator. The tested release is suspended and investigated; any revised configuration receives new versioning and independent testing.

A small sample may legitimately lead to AMBER. A noisy accuracy difference alone is not RED. **An AMBER or negative result still completes the research deliverable** — the programme is designed so that success is not defined after seeing the data.

12.5 Independent academic evaluation

A separate independent academic evaluation route is being developed within a national health system, on a design that would compare, for the same cases, the recommendation recorded by the treating MDT, the output of a version-locked Peregrine.Reason system assessing the same decision-time information, and an independent reference panel of thyroid specialists reviewing the same information.

The question would not be whether Peregrine reproduces the historical MDT decision — that decision is not treated as ground truth merely because a multidisciplinary group reached it. Nor is the expert reference infallible; it is an independently adjudicated comparator, and variation is not assumed to be error. Peregrine may perform better, similarly or worse, or the comparison may reveal legitimate equipoise. Each is a valid outcome.

That route is at proposal stage. **No academic award, institutional approval, sponsor, site or data access is claimed or assumed**, and any such study would remain under the control of independent academic investigators, including its protocol, analysis, interpretation and publication of unfavourable findings.

13. Bias, and what can honestly be claimed about it

Peregrine.Reason does not claim to eliminate bias. No AI system can.

Single-model systems inherit and amplify the biases of one training distribution, one model family, one institutional context and one prompting strategy. The architecture is designed to reduce that dependence and to make the residual bias visible, contestable and auditable.

Relevant sources include training-data bias, institutional practice bias, guideline bias, population-level evidence bias, demographic under-representation, specialty-specific bias, automation bias, majority-opinion bias, moderator bias, and clinician over-reliance on system output.

The controls are: independent reasoning before peer exposure; diversity across the four panel axes; separation of agent reasoning, moderation, safety review and clinician decision; no majority-vote rule; consequence-graded minority protection; explicit uncertainty and missing-data states; interpretable score vectors; sealed safety moderation; human-in-the-loop review; audit logs; version control; and post-deployment monitoring.

Evaluation examines not only whether outputs are clinically plausible, but whether particular clinical scenarios, case types or minority pathways are systematically underweighted. Where lawfully and ethically available, validation material varies across age, sex, comorbidity burden, disease stage, borderline findings, atypical presentations, incomplete data and uncertain guideline application — not to make demographic categories deterministic, but to detect uneven performance.

The correct claim is therefore not “*Peregrine.Reason is unbiased.*” It is: *Peregrine.Reason is designed to reduce single-model dependence by exposing disagreement, preserving minority signals, auditing reasoning, measuring correlated error, and keeping clinicians in control.*

14. Data protection

14.1 Phase I posture

Phase I is synthetic-first. The default input is synthetic or appropriately licensed curated material. Live patient data are not required to complete the funded build, and live record integration is outside Phase I scope.

This is stated plainly because it changes how the rest of this section should be read: the controls below are **design requirements for later clinical deployment**, developed and independently tested during Phase I against synthetic inputs, not a description of patient data currently in flight.

14.2 Requirements for any later clinical data route

Any route involving patient records requires an explicit controller and processor map, lawful basis, the necessary special-category conditions, purpose limitation, minimisation, approvals, an approved transfer mechanism, access controls, retention and deletion policy, and a supported processing environment.

Pseudonymisation does not automatically establish anonymity or remove governance obligations. The project will not base cross-border processing on a broad reading of a single court judgment or on a supplier’s marketing statement.

Andorran development must address the applicable Andorran data-protection framework, including Llei 29/2021. EU and UK data or deployments may introduce further jurisdiction-specific duties. An adequacy decision addresses a transfer mechanism; it is not a substitute for authority to process patient data, and it is not permission for unrestricted onward access.

Direct-identifier redaction is not the binding constraint in oncology — **quasi-identifiers are**. A rare histology combined with an age band and a region can single out an individual in a small population, and the populations this system is designed for are small.

The control that addresses this is **uniqueness screening**, not name redaction. Before any case leaves a controlled boundary, its attribute combination is tested against the local cohort distribution; combinations that are unique or near-unique are further generalised or withheld. Free text is the highest-risk surface and is either heavily structured before release or withheld entirely. Even structured de-identification can remove clinical context, so whatever is removed from a case is itself recorded, so that a performance difference can be attributed to input degradation rather than misread as a difference in reasoning quality.

Deployment resilience does not depend on any one cross-border transfer mechanism. If a legal basis, processor arrangement or permitted endpoint becomes unavailable, the routing policy blocks that route and uses an already-approved local or sovereign configuration (Section 3.9), or suspends the run. A substitute route requires its own legal and security assessment; contractual clauses are not treated as an automatic cure. Results and costs are reportable with and without external arms, so the consequence of losing that dependency is measurable rather than speculative.

14.3 Security design

Proposed controls include least privilege, tenant and environment separation, encryption in transit and at rest, protected secrets, scoped and short-lived credentials, strict egress restriction, dual independent de-identification checks where de-identification is used, and a separate identity-mapping service holding no clinical content.

Authentication to external providers must match the interfaces those providers actually support; no universal assertion is made that every vendor supports the same authentication mode. Unapproved egress is blocked.

Security testing addresses the risks specific to multi-agent language-model systems: prompt injection, contaminated retrieved passages, false case facts, tool misuse, malicious instructions embedded in documents, supply-chain change, excessive retention and unauthorised access. Unsafe or corrupted content can propagate through challenge, moderation and synthesis if it is not controlled at the boundary. **No special privilege or clinical action is granted because a model asks for it.**

A separate assessor, independent of the implementers, tests the agreed privacy and security scope before any patient-data route is enabled.

14.4 Trust-zone architecture and permissioned re-identification

Design for later clinical deployment. Not exercised on patient data in Phase I.

The intended clinical architecture divides the system into two trust zones. Identifiable patient data exists only inside the trusted zone and never crosses the boundary. Everything that crosses is pseudonymous by construction, having passed through a sanitisation gateway that performs identifier redaction, date shifting with a per-patient offset, geographic generalisation and the uniqueness screening described above. Gateway output is validated by a second, independently implemented detector; **disagreement between the two blocks release and raises a review task** rather than defaulting to permit.

The governing principle is worth stating directly, because it is the part of the design that is genuinely unusual:

The reasoning artefact travels. The identity does not.

The brief is a complete clinical document. It can be reviewed, audited, second-opinioned, evaluated and archived without ever resolving to a named patient. Resolution is a separate act, not a property of the document.

That act is governed by a **key escrow service**, held physically and logically apart from the case store, inside the trusted zone:

- It holds the pseudonym-to-identity mapping and **no clinical content**.
- Access is role-based and patient-scoped: only clinicians in the documented care chain for that specific patient may resolve their pseudonym.
- Resolution is explicit and logged, recording requester identity, timestamp and stated purpose. **There is no bulk export path.**
- Break-glass access for clinical emergency is permitted, but generates immediate notification and mandatory retrospective review.
- The escrow service is never reachable from any external reasoning path.

Two consequences follow that are easy to overlook and are recorded here rather than discovered later.

Externally hosted panel members necessarily see a slightly impoverished case relative to locally executed members. This is an accepted asymmetry, not a defect — but it must be treated as a covariate in evaluation, because a performance gap between arms may reflect input degradation rather than model capability.

Egress control is a layered scheme, not a single secret. Strong service authentication with short-lived scoped credentials rather than long-lived static keys; secrets held in protected storage and rotated; a strict outbound allowlist so the gateway can reach approved endpoints and nothing else, with all other egress denied at the network boundary; and every outbound payload hashed and recorded, so the exact content sent externally is reconstructible after the fact. Authentication modes must match what each provider actually supports, so no universal claim is made that one scheme applies everywhere.

14.5 Training

Clinical input is not used to train or fine-tune models by default. Any training use would require explicit governance approval, a lawful basis, documented handling, a de-identification or anonymisation strategy, institutional approval where required, auditability and separation from routine reasoning sessions. The default position is: **reasoning sessions are not training sessions.**

15. Clinical safety governance

Peregrine.Reason is governed as a clinical safety system, not merely as an AI interface.

Quality and clinical safety work begins during development rather than at submission: requirements traceability, hazard identification and logging, risk mitigation documentation, version control, change-impact review, defect management, output limitation rules, audit trails, incident reporting pathways and documented release decisions.

Safety governance applies across the agent layer, challenge layer, moderator, safety reviewer, clinician interface and audit workflow. The moderator is treated as safety-critical because it determines how uncertainty and disagreement reach the clinician.

Release controls and stopping rules. Gates are only meaningful if there is a defined response when one fails. Release requires an explicit, recorded decision on failed criteria, residual hazards and intended use — not the absence of an objection.

- A **single safety-critical omission** initiates case review and root-cause analysis.
- A **predefined sentinel failure, or repeated systematic failure**, suspends the affected release pending correction and re-testing.
- These are complementary levels of response, not competing stopping rules: the first is investigative, the second is suspensive.
- **No clinical output is released merely because an iteration budget is exhausted.** Running out of compute, time or revision cycles is a technical state, and a technical state is never converted into a clinical conclusion (T11, T14).

Certification, conformity assessment, clinical investigation permissions and external clinical effectiveness claims are later, separately gated activities.

16. Regulatory posture

At the current stage the system is described as **clinical reasoning support under development**.

It is not described as approved diagnostic AI, an autonomous tumour board, a treatment recommendation engine, a replacement MDT, an automated clinical decision-maker, or a clinically validated autonomous system.

Several positions are stated deliberately and should not be over-read:

- A **“research prototype” label does not itself determine medical-device status**. Intended purpose, functions, claims, users and target jurisdictions require a documented regulatory assessment.
- EU medical-device software qualification and classification guidance covers decision-support software, and **human sign-off does not automatically remove those obligations**.
- **No device class is pre-declared**. This document does not assign Peregrine.Reason to Rule 11 Class IIa, IIb or any other classification. That is an assessment to be made and documented, not asserted in advance.
- **Andorra is not treated as an EU member state**. EU obligations are assessed where the system is placed on the EU market, put into service, or otherwise falls within the relevant scope; the AI Act also addresses some third-country providers whose outputs are used in the Union. UK and Andorran routes require their own review.

- **Research exemptions are assessed precisely**, not assumed. Not every research-labelled deployment is exempt, and real-world testing conditions have their own requirements.

Regulatory posture is revisited as the system moves from research prototype to formative clinician evaluation, independent academic evaluation, and any later controlled deployment.

17. Innovation position and prior art

This programme does not claim to have invented multi-agent debate, argument graphs, judge models, abstention, multi-criteria dominance or evidence-to-decision reasoning. Each has research precedent, and the honest position is to name them.

Prior work	Established overlap	The question Peregrine still has to answer
Self-Consistency (Wang et al., ICLR 2023)	Samples reasoning paths and aggregates to a consistent answer	Whether frequency-independent safety escalation and preserved pathway sets improve uncertainty-rich clinical decisions
Multiagent Debate (Du et al., ICML 2024)	Multiple model instances debate and revise	Whether clinically material dissent should remain <i>unresolved</i> rather than end in a common answer
LLM-as-a-Judge (Zheng et al., NeurIPS 2023)	Evaluator models; position, verbosity and self-enhancement bias	Whether explicit clinical rule traces and adversarial controls reduce opaque judgment
MDAgents (Kim et al., NeurIPS 2024)	Adaptive medical collaboration, moderator review, external knowledge	Whether outcome-separated trade-offs, missing-data gates and decision-state design add measurable value
MedLA (Ma et al., AAAI 2026)	Logical trees, premise-level alignment, graph-guided discussion toward consensus	Whether an auditable graph can preserve legitimate equipoise and safety objections without forced ranking
MediQ (Li et al., NeurIPS 2024)	Missing-information elicitation and abstention in interactive clinical reasoning	Whether decision pivots and conditional pathways improve MDT preparation in the bounded domain
Multi-agent peer-reviewed reasoning (Zhan et al., 2026)	Independent reasoning followed by peer quality review rather than answer agreement	Whether the clinical pathway-set, proportionality and safety architecture improves on selecting a highest-rated reasoning chain

Decision-science antecedents are equally explicit. The comparison rule uses Pareto-style dominance from the established multi-criteria decision analysis family, and GRADE Evidence-to-Decision frameworks already structure evidence, patient values, benefits, harms and contextual judgments.

The distinctive design under test is a specific integration, not any single component: supported pathway sets and structured equipoise; frequency-independent, consequence-sensitive escalation; outcome-separated, null-safe benefit–harm objects; an auditable admissibility and dominance policy; and a sealed two-pass safety review over the complete pre-filter claim register.

Whether that integration provides measurable value beyond the stated comparators is an open experimental question. No claim of world-first status, exhaustive prior-art clearance, freedom to operate or patentability is made.

17.1 The unresolved technical questions

These are the problems the programme exists to work on, and none of them is solved:

- extracting *material* rather than rhetorical claims from model output;
- separating absence of evidence from falsehood;
- keeping quantitative benefit and harm estimates genuinely comparable;
- handling uncertainty without making the system uselessly indecisive;
- preserving important dissent without alert flooding;
- maintaining safety behaviour when models or deployment services change underneath the system.

The outcome is uncertain, and a simpler architecture may prove preferable.

18. Development roadmap

The roadmap is milestone-based. Phase I is a self-contained twelve-month programme; subsequent phases depend on results, independent approvals and funding.

Phase I — research prototype and formative evaluation Requirements, decision ontology and Clinical Decision Policy Record; platform, model abstraction and privacy/security infrastructure; independent reasoning, challenge and revision; the Peregrine Moderator Engine; sealed two-pass safety review; thyroid demonstrator and clinician interface; independent case production, reference lock and formative comparison; release dossier and a GREEN/AMBER/RED progression decision.

Phase II — independent academic evaluation Independent evaluation of a version-locked release, under an independently approved protocol, with the sponsor controlling design, analysis, interpretation and publication.

Phase III — clinical governance preparation Data protection impact assessment; clinical safety case; cybersecurity review; model provenance framework; regulatory assessment against intended purpose and jurisdiction; human oversight process; deployment governance.

Phase IV — controlled deployment Institutionally governed pilot; monitored use; safety reporting; post-deployment review; clinician feedback; drift monitoring; expansion only after validation.

Each phase is a gate, not a schedule commitment. No phase authorises the next until its evidence supports it.

18.1 Integration with clinical systems

Beyond Phase I. Live electronic health record integration is outside the Phase I scope. Phase I runs on synthetic and curated case material ingested through structured export, and the formative comparison does not depend on a live record connection.

Integration is nevertheless a precondition for any real clinical use, and the architecture is designed so that it can be added without redesign. The intended direction is a **read-only** interface: standards-based retrieval of the case fields defined in the clinical schema — structured demographics, imaging and pathology report text, biochemistry, staging and risk classification, prior treatment — mapped through recognised clinical terminologies onto the case representation.

Three constraints are stated now because they shape the design rather than being retrofitted to it:

- **The platform never writes to the clinical record.** A reviewed brief is filed by the responsible clinician through normal channels, not by the system. Keeping Peregrine.Reason outside the record of truth simplifies both the regulatory position and an institution's own change-control process.
- **Integration scope is bounded to the defined case fields.** This is not a general-purpose interface, and uncontrolled scope growth here is the most likely way an integration becomes an unmanaged dependency.
- **Integration must not weaken data minimisation, auditability or clinical safety.** Retrieving more because retrieval is available is not an improvement.

The practical risks are institutional rather than technical: interface engine access, vendor professional-services arrangements and site IT scheduling are outside a supplier's control and routinely determine the timeline. **No integration is claimed, and none should be assumed, until it is implemented and validated at a named site.**

18.2 Long-term direction: a distributed reasoning network

Beyond Phase I. This section describes intended direction, not funded work, and nothing in it is a current capability.

The Phase I system runs a small panel within a single controlled environment. That is a deliberate feasibility choice for a programme whose research question is the moderation layer. It is not the end state.

The long-term architecture is a **distributed reasoning network**: independent nodes, each running an individual reasoning agent, participating in the same structured challenge and moderation protocol across institutional and organisational boundaries.

A node may represent a distinct model, a distinct reasoning framework, a specialty-specific agent, a hospital-controlled deployment, a validated clinical reasoning module, a privacy-preserving local deployment, or a governance-approved external reasoning service. Each node contributes structured claims, evidence, uncertainty and challenge. **No node controls the final clinical output**, and no node is an autonomous decision-maker.

Why distribution, rather than a larger panel in one place. The value is not scale. It is that the participants stop sharing a single training distribution, a single institutional assumption set, a single prompting strategy and a single operator. Correlated error is the failure mode that aggregation cannot detect (Section 8.3), and the strongest available defence against it is genuine heterogeneity of origin and control. A distributed network also allows institutions to contribute reasoning without surrendering control of their data or their models, which is a governance property, not a technical one.

Moderation in a distributed network. The same constraint applies with more force: the moderator must not become the new centralised authority. In the long term the moderator function may itself be modular, distributed, institutionally configurable or independently validated — provided clinical safety, auditability and output consistency are preserved, and provided different moderator implementations remain comparable. Peregrine.Reason should not depend permanently on one opaque moderator as the single arbiter of reasoning quality.

Decentralised in reasoning, governed in data access. Distribution of reasoning does not imply distribution of patient data, and the two must not be conflated. Patient data is not sent to arbitrary open nodes. Any clinical participation requires controlled node admission, security standards, privacy controls, auditability, institutional oversight, data minimisation and explicit governance of data flow.

Node admission would require, at minimum, a documented profile covering model identity and version, deployment environment, known limitations, evaluation history, update history, permitted use cases, data retention behaviour and safety constraints — with output traceable to the node and version that produced it.

What has to be true first. This direction is contingent, and the contingencies are substantial: Phase I must demonstrate that structured moderation earns its complexity at all; node admission, liability and clinical governance across institutional boundaries are unsolved organisational problems, not engineering ones; and distributed operation multiplies the attack surface, the audit burden and the regulatory question. None of these is resolved here, and if Phase I shows that a simpler architecture performs as well, the distributed network is a direction that should be abandoned rather than pursued.

19. Openness

Peregrine.Reason moves toward openness where openness improves trust.

Components suited to publication include the reasoning protocol, challenge schema, moderator evaluation framework, output templates, audit methodology, safety and validation methodology, and scoring rubric design — the parts that allow others to check whether the claims in this document are true.

Components that remain controlled include patient data, clinical datasets, production credentials, security-sensitive infrastructure, institution-specific configurations, regulated deployment logic, and any component where openness would weaken privacy or safety.

The principle is: **open where trust requires transparency; protected where privacy and safety require control.**

20. Positioning

Peregrine.Reason is:

- not a chatbot;
- not diagnostic AI;
- not a guideline engine;
- not an autonomous MDT replacement;
- not a black-box recommendation system;
- not a system that replaces clinician judgment.

It is a clinical reasoning architecture designed to make evidence, uncertainty, competing pathways and minority safety concerns inspectable — decentralised in reasoning, structured in moderation, safety-checked by design, and clinician-led in decision-making.

Its purpose is not to produce one answer. Its purpose is to make the reasoning behind a decision transparent, plural, auditable and clinically useful — and to be testable enough that it can be shown not to work.

Most AI systems optimise for agreement. Peregrine.Reason preserves disagreement where disagreement is material, and says so when it cannot tell.

The clinician remains the decision-maker.

21. Selected references

Clinical

Clinical sources are cited by publication record — title, journal, volume, pages and DOI — which identifies each item precisely and resolves to the full author list at the publisher.

6. 2025 American Thyroid Association Management Guidelines for Adult Patients with Differentiated Thyroid Cancer. *Thyroid*. 2025;35(8):841–985. doi:10.1177/10507256251363120
7. Optimization of the risk-benefit ratio of differentiated thyroid cancer treatment. *Eur J Surg Oncol*. 2018;44(3):276–285. doi:10.1016/j.ejso.2018.01.077
8. Aggressive differentiated thyroid cancer. *Eur J Surg Oncol*. 2018;44(3):367–377. doi:10.1016/j.ejso.2017.09.019
9. A study of variation in therapeutic approach to low-risk differentiated thyroid cancer in the UK. *J Laryngol Otol*. 2024;138(1):83–88. doi:10.1017/S0022215123000841
10. Not all guidelines are created equal: rigour, evidence, and the case for transparent adaptation in papillary thyroid cancer guidance [editorial]. *Eur J Surg Oncol*. 2026;52(7):111907. doi:10.1016/j.ejso.2026.111907
11. Assessment of the Scandinavian guidelines on papillary thyroid cancer using the AGREE II instrument. *Eur J Surg Oncol*. 2026:111876. doi:10.1016/j.ejso.2026.111876
12. Thyroid cancer surgery guidelines in an era of de-escalation. *Eur J Surg Oncol*. 2018;44(3):297–306. doi:10.1016/j.ejso.2017.03.005

Conceptual antecedents

These are acknowledged as intellectual background for the design argument in Section 3, not as evidence for any claim about the system.

8. Hayek FA. The Use of Knowledge in Society. *American Economic Review*. 1945;35(4):519–530. — distributed knowledge and the limits of central planning.
9. Mill JS. *On Liberty*. 1859. — the value of dissent, and the suppression of minority opinion.
10. de Tocqueville A. *De la démocratie en Amérique*. 1835–1840. — the tyranny of the majority.
11. Shaw GB. *Maxims for Revolutionists*. 1903. — progress and the refusal to adapt to prevailing opinion.

Methodological

12. Cochrane Handbook for Systematic Reviews of Interventions, Chapter 15: Interpreting results and drawing conclusions.
13. Thokala P, Devlin N, Marsh K, et al. Multiple Criteria Decision Analysis for Health Care Decision Making — an introduction. ISPOR MCDA Emerging Good Practices Task Force, Report 1. *Value Health*. 2016;19(1):1–13.
14. Alonso-Coello P, Schünemann HJ, Moher J, et al. GRADE Evidence to Decision (EtD) frameworks. *BMJ*. 2016;353:i2016.

AI and multi-agent antecedents

15. Wang X, Wei J, Schuurmans D, et al. Self-Consistency Improves Chain of Thought Reasoning in Language Models. ICLR 2023.

16. Du Y, Li S, Torralba A, Tenenbaum JB, Mordatch I. Improving Factuality and Reasoning in Language Models through Multiagent Debate. ICML 2024.
17. Zheng L, Chiang WL, Sheng Y, et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. NeurIPS 2023.
18. Kim Y, Park C, Jeong H, et al. MDAgents: An Adaptive Collaboration of LLMs for Medical Decision-Making. NeurIPS 2024.
19. Ma S, Huang J, Zhang F, et al. MedLA: A Logic-Driven Multi-Agent Framework for Complex Medical Reasoning with Large Language Models. AAAI 2026.
20. Li SS, Balachandran V, Feng S, et al. MediQ: Question-Asking LLMs and a Benchmark for Reliable Interactive Clinical Reasoning. NeurIPS 2024.
21. Zhan Z, Zhou S, Zhang R. Let large language models judge each other: multi-agent peer-reviewed reasoning for medical question answering. *J Am Med Inform Assoc.* 2026.

Regulatory and security

22. Medical Device Coordination Group. MDCG 2019-11 rev.1: Guidance on Qualification and Classification of Software under Regulations (EU) 2017/745 and 2017/746.
23. Regulation (EU) 2024/1689 (Artificial Intelligence Act).
24. Llei 29/2021, del 28 d'octubre, qualificada de protecció de dades personals (Andorra).
25. OWASP Top 10 for Large Language Model Applications.
26. IMDRF. Software as a Medical Device: Clinical Evaluation.

Published by the Peregrine.Reason programme, Andorra.

Peregrine.Reason — www.peregrinereason.ai

This document describes research and development in progress. It is not a claim of clinical performance, regulatory status or commercial availability.