AI Epidemiology: achieving explainable AI through expert oversight patterns

Kit Tempest-Walters
ktempestwalters@gmail.com

Abstract:
AI Epidemiology is a framework for governing and explaining advanced AI systems by applying population-level surveillance methods to AI outputs. The approach mirrors the way in which epidemiologists enable public health interventions through statistical evidence before molecular mechanisms are understood. This bypasses the problem of model complexity which plagues current interpretability methods (such as SHAP and mechanistic interpretability) at the scale of deployed models.

AI Epidemiology achieves this population-level surveillance by standardising capture of AI-expert interactions into structured assessment fields: risk level, alignment score, and accuracy score. These function as exposure variables which predict output failure through statistical associations, much like cholesterol and blood pressure act as exposure variables predicting cardiac events. Output-failure associations are subsequently validated against expert overrides and real-world outcomes.

The framework places zero burden on experts and provides automatic audit trails by passively tracking expert convergence and divergence with AI recommendations. Since it analyses outputs rather than internal model computations, it also provides governance continuity when institutions update models and switch vendors. Finally, by providing reliability scores and semantic assessments (e.g. 'this recommendation resembles 500 cases overridden by experts due to guideline violations'), it enables experts and institutions to detect unreliable AI outputs before they cause harm. This democratises AI oversight by enabling domain experts to govern AI systems without requiring machine learning expertise.

1 Introduction

Modern AI systems present a fundamental governance challenge. We deploy increasingly powerful models in high-stakes domains such as healthcare, finance, and criminal justice, but struggle to explain their internal computations in human-understandable terms. This opacity

becomes critical when such models produce outputs that are dangerous, unethical, and inaccurate. There is therefore an urgent need to achieve governance-oriented explanation of large systems, which has thus far eluded interpretability methods.

We assign the term ‘correspondence-based interpretability’ to those methods which seek to establish correspondence between what the model focuses on internally and what it outputs. This includes mechanistic interpretability, which attempts to trace computational circuits and attention patterns, feature attribution methods like SHAP (Lundberg and Lee 2017) and LIME (Ribeiro et al. 2016) that score which input features the model weighs most heavily, and even chain-of-thought prompting that assumes the model's stated reasoning faithfully represents its actual computational process (Turpin et al. 2023). All these methods share the fundamental assumption that we can identify what the model is ‘paying attention to’ and map this focus to its outputs in human-understandable terms. However, this correspondence-based paradigm becomes computationally intractable and epistemically unreliable as models scale. These scaling challenges manifest as features interacting in more complex ways, attributions becoming unstable across similar inputs, and models generating plausible-sounding explanations that don't reflect their actual decision processes. While establishing reliable correspondence between internal model computations and human concepts remains a worthy scientific goal, these methods cannot provide the robust governance we need for deployed systems today.

In the light of this challenge, this paper proposes AI Epidemiology as a framework that bypasses the intractable problem of correspondence-based interpretability by applying population-level surveillance methods from public health. It does this by identifying and mitigating AI risks through systematic observation of outputs and expert interventions without requiring transparency into model computations, similar to the way in which epidemiological methods have enabled public health interventions without requiring mechanistic understanding. The framework asks which outputs experience failure, what observable characteristics predict these failures, and where failures concentrate across domains, jurisdictions, and models.

AI Epidemiology therefore pursues the goal of systematising human oversight that is already interpretable rather than attempting to make opaque models transparent. Since domain experts in medicine, finance, and law routinely record their decisions electronically, these decisions can be collated to create automated audit trails and reveal patterns of expert oversight of AI recommendations. For instance, if oncologists consistently override certain treatment suggestions or loan officers regularly reject specific credit profiles, systematic patterns emerge

which can aid in future oversight decisions. Furthermore, this expert oversight can be explained in reference to institutional guidelines and validated against outcomes. We can therefore identify which output characteristics require oversight by systematically capturing expert interventions and analysing patterns at the population level, leading to AI explainability and governance at scale.

The following sections develop this framework in detail. Section II situates AI Epidemiology within the broader landscape of interpretability research and statistical monitoring. Section III operationalises the approach through the Logia protocol and its four core components. Section IV presents empirical results from a feasibility study in ophthalmology. Section V compares the explanatory power of correspondence-based methods and AI Epidemiology using a chess demonstration. And Section VI addresses methodological and implementation challenges.

2 Beyond correspondence: situating AI Epidemiology

Having touched upon the limitations of correspondence-based interpretability, we now examine how AI Epidemiology offers an alternative pathway grounded in historical precedent. In the tradition of John Snow identifying contaminated water sources without knowing the bacterial pathogen, and Bradford Hill linking smoking to lung cancer without understanding the molecular mechanisms, AI Epidemiology seeks to identify and intervene in AI failure patterns without requiring mechanistic transparency into model computations. This reframing has profound implications for AI governance. Mechanistic interpretability research pursues the worthy goal of understanding how neural networks compute, analogous to the quest of molecular biology to explain disease through cellular mechanisms. But as we acknowledge that public health should not rely solely on biological understanding, so too must AI explainability proceed without complete mechanistic understanding. AI Epidemiology enables urgent action by replacing attempts to decode model internals with systematic pattern recognition from expert interventions. This approach democratises AI explainability since domain experts in fields such as medicine, law, and finance can contribute to and understand AI explanation and oversight without requiring expertise in computer science or machine learning.

The Bradford Hill precedent

Austin Bradford Hill and Richard Doll demonstrated that systematic epidemiological evidence could justify public health action without mechanistic understanding by establishing the causal link between smoking and lung cancer. They accumulated overwhelming statistical evidence

through two landmark studies: the 1950 case-control study comparing 709 lung cancer patients with controls (Doll and Hill 1950), and the British Doctors Study launched in 1951 following 34,439 male physicians over five decades (Doll and Hill 1956; Doll et al. 2004). Results from these and other studies demonstrated a nine- to thirty-fold increased risk of lung cancer among smokers, clear dose-response relationships (heavier smoking predicted higher cancer rates), and temporal precedence (smoking preceded cancer diagnosis) (Hill 1965).

On the basis of this accumulating evidence, major public health actions proceeded: the US Surgeon General declared smoking a cause of lung cancer in 1957 (Burney 1959), as did the Royal College of Physicians of London (Royal College of Physicians, 1962), and the landmark 1964 US Surgeon General's Advisory Committee report reviewed over 7,000 scientific articles to conclude definitively that smoking causes lung cancer and chronic bronchitis (U.S. Department of Health, Education, and Welfare, 1964). Federal regulation followed with the 1965 Cigarette Labeling and Advertising Act, which mandated warning labels on cigarette packaging, and the 1970 Public Health Cigarette Smoking Act, which banned cigarette advertising on television and radio. These interventions prevented an estimated eight million deaths in the United States alone between 1964 and 2014 (Holford et al. 2014).

However, the specific molecular mechanisms by which tobacco smoke causes cancer remained unknown during this entire period of public health action. BPDE-DNA adducts were identified only in 1976 (Weinstein et al., 1976), and detailed mapping of p53 tumour suppressor gene mutations in smokers' lung tissue emerged in 2002 (Pfeifer et al. 2002), decades after the causal determination and policy implementation. Hill's key contribution was to codify the way in which statistical and epidemiological evidence establishes causality. In his 1965 paper, Hill proposed nine ways of inferring causation from correlation, and noted that focusing primarily on essential criteria such as temporal precedence and dose-response relationships could save lives while biological research was still ongoing (Hill 1965).

This methodological principle translates directly to AI Epidemiology, which assesses the statistical risk of AI outputs being inaccurate or misaligned.

From correspondence to risk stratification: an epistemic shift

The transition from correspondence-based interpretability to AI Epidemiology represents a fundamental epistemic shift in how we approach AI explanation and oversight. In order to

understand this shift, we need to define what each paradigm seeks, how it validates knowledge claims, and the form that explanation takes.

The correspondence model seeks to make AI models transparent by producing faithful explanations of what the model is focusing on when producing its predictions. For example, SHAP (SHapley Additive exPlanations) tests different combinations of words across model input features to identify the extent to which each contributes to the final output. However, it is computationally impossible to test all possible combinations of words in an LLM containing thousands of tokens, which forces researchers to resort to approximations which can be unstable and misleading (Horovicz and Goldshmidt 2024). Moreover, recent work shows that adversaries with access to the model can manipulate SHAP explanations without altering the model's outputs. One attack achieved this by manipulating SHAP into drastically undervaluing a sensitive feature's role in the outcome through biased sampling of reference data (Laberge et al. 2023). In another experiment, a fairness audit was blinded to bias after an attack reduced the apparent importance of a race input by 90% (Laberge et al. 2023). As a result, the SHAP-based audit reported far less bias even though the model's outputs were just as discriminatory as before.

Furthermore, observational methods in mechanistic interpretability such as attention visualisation, feature visualisation, and probing also struggle with combinatorial complexity and superposition in Transformer models (Olah et al. 2020). This is because individual components such as neurons and attention heads often encode multiple unrelated human concepts, thereby undermining the ideal of a 'grandmother neuron' for each concept. As Olah et al. (2020) and others have demonstrated, neurons are frequently polysemantic since they activate in response to a variety of features, e.g. cat faces, fronts of cars, and cat legs. This directly challenges a key assumption of observational mechanistic interpretability that internal model components correspond to discrete, intelligible functions that can be analysed in isolation.

Interventional mechanistic interpretability has attempted to overcome some of these observational limitations by testing which features causally drive specific behaviours rather than mapping the correlation between them (Bereska and Gavves 2024). Techniques such as causal ablation, causal tracing, and model editing investigate how model behaviour changes when components such as neurons, attention heads, or weights are altered. For example, researchers ablate specific attention heads by setting their outputs to zero to test whether model

performance degrades. If performance degrades, this suggests that the head was causally important for the task. Similarly, localised model editing tweaks individual neuron weights to test whether this causes behavioural changes, thereby enabling researchers to attribute certain behaviours and concepts to specific neurons.

While these methods have achieved notable successes in small-scale settings[1], they also face fundamental problems of scale and entanglement. In large Transformer models, behaviours are distributed redundantly rather than concentrated at single points of failure (He et al., 2024). For instance, ablating an attention head can lead to other components compensating for it, which conceals the ablated component's contribution. Such an intervention might not have a measurable effect on model performance, which can lead researchers to conclude incorrectly that the original component was unimportant. Alternative interventional methods such as causal tracing, which resamples or patches activations to locate where specific information is stored, also show that information is entangled across many neurons and layers rather than being cleanly localised.

These examples are indicative of the difficulties that correspondence-based methods face in pursuing introspective faithfulness in contemporary models. On the one hand, feature attribution methods face combinatorial complexity and manipulation vulnerabilities, and on the other, mechanistic interpretability encounters polysemanticity, entanglement, and compensation effects (McGrath et al. 2023). Since the most up-to-date models have billions or even trillions of parameters, comprehensive validation through introspective faithfulness remains a long-term research goal rather than a practical solution for urgent governance needs. Indeed, although isolated circuits have been identified in large models such as L3.1-8B (Nam et al.2025) and Claude 3.5 Haiku (Ameisen et al. 2025), the foundational correspondence-based question of how models compute their outputs remains largely unanswered at scale.

AI Epidemiology asks fundamentally different questions: Which outputs experience failure? What observable characteristics predict these failures? And where do failures concentrate across domains, jurisdictions, and models? Failure refers to an output being overridden by an expert or being accepted by an expert with an undesirable outcome. Observable characteristics that predict these failures include low accuracy and low alignment scores. For example, if an output

[1] For example, Bau et al. (2019) found that editing certain neurons in a translation model could reduce gender bias, and Meng et al. (2022) showed that modifying a single MLP-layer weight in GPT-2 XL could erase a specific model memory.

is incorrect and goes against ethical or procedural norms, this would be considered low accuracy and low alignment and would predict either expert disagreement or an undesirable outcome. Patterns would soon emerge to show which outputs tend to be unreliable and present the most risk to stakeholders. For example, AI outputs in medicine may have a far higher expert agreement rate than in law. In medicine, outputs recommending oncology treatments may have high accuracy and high alignment, whereas those recommending blood pressure treatments may be low accuracy and/or low alignment. Similarly, general practitioner disagreements may be far less common than specialist disagreements. These data points would in the first instance help experts to identify potentially harmful outputs and mitigate risk before acting on them. In addition, harmful outputs could be stratified for risk level based on the stakes of the case, previous expert decisions, and validated outcomes.

This reframes what we mean by an explanation. As precedent, an epidemiological explanation might convey that that a 60-year-old male smoker with elevated blood pressure and cholesterol has a 20 percent ten-year risk of cardiovascular disease. It does not tell us that the patient has atherosclerosis or convey specific information about his arteries or circulatory system. Instead, this explanation enables clinical decision-making including initiating preventive medications, intensifying blood pressure management, and counselling on smoking cessation, based on patterns in the data. Similarly, an epidemiological explanation of AI outputs explains that a model's oncology recommendation carries low reliability because it resembles a pattern of similar cases which contradicted treatment guidelines and were overridden 75% of the time by experts. This does not explain the relationship between internal model computations and human-understandable concepts, but is potentially of greater use in detecting risk while research into model transparency continues.

AI Epidemiology therefore necessitates a shift from understanding behaviour through internal processes to understanding behaviour through observable patterns and their consequences. Both are legitimate forms of explanation which serve different purposes on different timescales. While correspondence-based interpretability pursues the long-term scientific goal of understanding AI systems as computational objects, AI Epidemiology pursues the governance goal of identifying risk and enabling appropriate oversight at scale.

AI Epidemiology vs statistical monitoring

AI Epidemiology must also be distinguished from existing statistical monitoring systems for AI models. Statistical monitoring systems such as IBM Watson OpenScale and Google Vertex AI Model Monitoring measure model outputs in aggregate against outcomes for the purpose of model evaluation (Naveed et al. 2025). For example, a statistical monitoring system tracking a credit lending model might measure that over 10,000 recent loan decisions, the model achieved 91 percent accuracy when comparing implemented recommendations against actual outcomes. They therefore tell the institution that failures have occurred in aggregate but cannot explain which types of outputs have failed or why. As a result, statistical monitoring tools are unable to flag risk in advance of potentially dangerous and/or biased decisions.

AI Epidemiology addresses this gap by providing output-level risk assessment informed by population-level surveillance. For each credit application, the system evaluates the accuracy and alignment of the AI output based on observable characteristics corresponding to previous outputs in conjunction with expert overrides and outcomes. It then generates a semantic assessment by comparing the AI output to historical patterns, e.g., 'this AI output is unreliable because it resembles 500 prior recommendations overridden by experts due to inaccuracy, and accepting such recommendations led to a 20 percent higher loan default rate.' The loan officer therefore receives a specific and timely assessment of why a particular output is unreliable based on patterns from similar past cases, rather than a non-specific, post-hoc alert that the performance of the model has declined and requires recalibration.

The distinction parallels disease surveillance versus clinical epidemiology. Disease surveillance tracks incidence rates, e.g. 'influenza rates increased 40 percent this week in the Northeast region,' providing population-level monitoring analogous to statistical model monitoring. By contrast, clinical epidemiology enables individual patient risk stratification by applying population patterns to individual decision-making (Fletcher et al. 2020), e.g. 'this 65-year-old patient with diabetes and heart disease has a 15 percent risk of severe influenza complications, warranting early antiviral treatment'. Disease surveillance and clinical epidemiology therefore serve complementary purposes: the former detects trends and triggers systemic responses, while the latter enables targeted intervention, allocating intensive resources to high-risk patients and providing standard care for low-risk patients.

AI Epidemiology brings this clinical epidemiology capability to AI oversight by providing output-level risk assessment, complementing the role of statistical monitoring in tracking aggregate model performance.

3 Core components: operationalising AI Epidemiology

AI Epidemiology is operationalised through three integrated components and a dual stratification system, each answering different epidemiological questions and together forming a complete framework for output-level AI oversight. The three operational components are: (1) *Logia Grammar* for structured exposure assessment, (2) Expert action and outcome validation for capturing and validating interventions, and (3) *Tracelayer* for pattern analysis and model correction. These components generate dual stratification through (4) assessment of consequence severity and of output failure probability, enabling risk-appropriate resource allocation. This framework is applicable across different domains and AI models.

Logia Grammar: standardised measurement of AI outputs

AI Epidemiology requires a structured protocol for systematically categorising AI outputs to enable semantic explanation. The Logia Grammar provides this structure by capturing outputs in a standardised format that enables population-level pattern recognition while explaining individual cases through comparison to these patterns. Every model interaction begins with three input-output components: *mission* (the user's query or instruction), *conclusion* (the model's answer), and *justification* (the model's reasoning or evidence). These capture what the model was asked and how it responded. The grammar then adds three assessment fields: *risk level* (high, medium, or low assessment of potential harm), *alignment score* (degree of correspondence to institutional guidelines and practices), and *accuracy score* (correctness of factual claims and reasoning). These assessment fields represent structured review of the AI's output. Finally, the grammar captures two expert action fields: *override* (yes/no indicator of whether the expert's actual decision deviated from the AI recommendation) and *corrective option* (the alternative action the expert took when override occurred). The expert action fields document expert decisions taken in response to the AI output.

The basic grammar comprises the input-output fields (mission, conclusion, justification) and the two expert action fields (override, corrective option), as these allow institutions to track expert-AI convergence and divergence. Together, these five fields create a complete record of each expert-

AI interaction: the AI suggested X based on reasoning Y, and the expert either accepted this (override: no) or chose alternative Z (override: yes, corrective option: Z). This basic structure provides institutions with automated audit trails by documenting every expert-AI interaction, including those in which human judgment diverged from algorithmic suggestions.

Critically, all fields are populated through passive background monitoring without requiring any manual data entry from experts: mission, conclusion, and justification are automatically extracted from the user's original input and the source model's output, thereby standardising different conversational formats into comparable structured elements. These fields are automatically updated as the interaction progresses (for instance, if the AI's conclusion changes in response to additional information or through multi-turn conversation). The expert action fields (override and corrective option) are captured automatically through middleware integration with institutional electronic systems. The system therefore monitors whether the expert's actual decision in their workflow system matches the AI's recommendation. For example: if mission is 'treat patient condition X,' conclusion is 'prescribe medication Y,' but the expert actually prescribes medication Z in their electronic health record, the system automatically records override: yes, and corrective option: prescribe medication Z. This enables experts to work normally in their existing systems while oversight documentation is automatically collated in the background.

While the basic grammar enables a foundational audit trail by documenting expert-AI convergence and divergence, the assessment fields substantially enrich this audit capability while simultaneously enabling explainability and correction of AI outputs. The risk level assesses case severity based on the mission field, indicating the stakes involved in the decision; the alignment score evaluates whether the conclusion and justification conform to institutional guidelines and practices; and the accuracy score measures the factual correctness of claims and reasoning in the AI's output. Experts neither fill out any fields, nor see the underlying scores. This invisibility preserves data integrity by ensuring expert decisions represent genuine revealed preferences rather than responses to visible scores, which would introduce observer effects (McCambridge et al. 2014) and compromise the supervised learning process.

Initially, retrieval-augmented generation (RAG; Lewis et al., 2020) produces each assessment by comparing AI outputs to specific institutional documents. However, these scores are dynamically calibrated over time through supervised learning (Goodfellow et al. 2016) using a triple-signal

system: RAG provides rapid initial assessment, expert decisions (overrides and acceptances) provide medium-term validation, and tracked outcomes provide the most reliable, longer-term signals. For instance, RAG might initially score an output as having high accuracy because it closely matches documented facts and guidelines. However, when multiple experts override similar outputs and provide corrective options, this prompts recalibration to medium accuracy. Subsequently, when outcome tracking confirms these overrides prevented mistakes (e.g., incorrect diagnoses or treatment recommendations), the model learns to classify such outputs as having low accuracy. This progression reflects increasing precision: RAG measures document compliance, expert decisions reveal errors that require professional judgment to detect, and outcomes provide empirical validation.

Furthermore, the alignment and accuracy fields function as exposure variables as opposed to simple descriptive categories. Whereas descriptive categories label what has already occurred for the purpose of post-hoc analysis, exposure variables are measurable characteristics that predict future outcomes through statistical associations (Lee and Pickard 2013). In epidemiology, blood pressure readings are exposure variables because population data reveals that elevated readings predict future cardiovascular events with quantifiable probability. For example, in 1948, the Framingham Heart Study established through statistical analysis of over 5,000 participants that measurable characteristics such as blood pressure and cholesterol predict cardiovascular events (Dawber et al., 1951). Similarly, Logia's risk level, alignment score, and accuracy score function as exposure variables by accumulating associations between observable output characteristics and expert intervention patterns across thousands of cases. Thus, when outputs with low alignment and citations to preliminary research are overridden in 78% of cases, this denotes the discovery of a predictive risk factor enabling prospective triage. The system can then flag similar outputs for mandatory review before they are acted upon, which prevents potential harm through statistical association rather than mechanistic understanding of how the AI model computes its outputs.

The Logia Grammar therefore structures AI-expert interactions into fields that enable analysis over thousands of cases. The input/output fields record AI recommendations and expert responses, while assessment fields function as exposure variables that predict intervention risk. This provides explainable and scalable oversight by enabling institutions to identify and prevent problematic outputs before they cause harm.

Finally, these benefits are model-agnostic since semantic patterns are tracked in aggregate through the Logia Grammar fields, and flagged for misalignment or inaccuracy independent of model considerations. Institutions then receive both model-agnostic output assessments e.g. '75% of such outputs were misaligned' and model-specific output assessments e.g. '58% of such GPT-4 outputs were misaligned'. As new models are introduced, model-agnostic output assessments become the default until enough data accrues about the new models. This allows institutions to use updated models, run A/B tests, and switch vendors without losing governance or explainability functionality. By contrast, correspondence-based methods such as SHAP and mechanistic interpretability must analyse each model in isolation.

Expert action and outcome validation

An outcome variable is the event or condition that exposure variables seek to predict through statistical association (Lee and Pickard 2013). In AI Epidemiology, expert override and adverse outcomes are the outcome variables, because these are the events which the exposure variables (alignment score and accuracy score) are designed to predict. The system tracks override rates across these two exposure profiles to identify which AI output characteristics correlate with expert intervention. As patterns emerge which show that outputs with specific characteristics consistently generate overrides and adverse outcomes, the system produces a *reliability score* and explanation for expert review before a decision is made. This enables pre-hoc intervention rather than post-hoc correction.

The exposure-outcome structure parallels traditional epidemiology. Epidemiologists map statistical correlations between exposure variables (symptoms, behaviours, biomarkers) and disease incidence to enable clinical intervention (Lee and Pickard 2013). Similarly, AI Epidemiology maps statistical correlations between AI output exposure variables (alignment, accuracy) and AI output 'disease incidence' (as represented by expert override) to enable governance intervention. In standard epidemiology, strong statistical correlation between an exposure variable and an outcome variable are sufficient grounds for judgment. For example, when the Framingham Heart Study established that elevated blood pressure predicts cardiac events, and smoking studies demonstrated that tobacco exposure predicts lung cancer, these exposure-outcome pairings alone justified clinical action. However, AI Epidemiology present more complex exposure-outcome relationships because the first outcome variable is expert override, which represents human judgment rather than real world outcomes.

This makes exposure-outcome analysis in AI Epidemiology riskier than in public health due to the possibility of expert bias becoming entrenched. For example, loan officers might impose stricter lending conditions on certain demographic groups despite equivalent default risk (Brock and De Haas 2021); doctors might override treatment recommendations based on practice patterns unsupported by evidence (Alexander 2019); and lawyers might reject favourable settlement offers based on overconfidence bias rather than actual trial prospects (Kiser et al 2008). A reliability assessment for AI outputs that depends solely on such expert overrides could validate these errors by treating expert consensus as ground truth, leading to dogmatic rather than empirical justification. AI Epidemiology mitigates this risk through real-world outcome tracking (the second outcome variable) which validates whether expert interventions actually improve outcomes.

On the one hand, risk and accuracy are validated through empirical outcomes which reveal whether expert interventions prevented adverse events (validating risk assessments) or proved correct (validating accuracy assessments). Those outcomes that confirm expert interventions prevented harm reinforce existing risk classifications. Conversely, cases flagged as high risk or low accuracy that experienced no adverse events after expert acceptance reveal that the system's criteria require recalibration. In this way, empirical validation ensures that risk and accuracy assessments reflect actual consequences rather than perceived threats.

On the other hand, alignment is validated through procedural outcomes. Since the alignment score measures adherence to institutional standards rather than empirical correctness, it cannot be validated by whether decisions produced accurate outcomes. Instead, the system tracks regulatory violations, professional standards breaches, and stakeholder complaints. Similar to risk and accuracy, cases that experts reject despite high initial alignment scores and that subsequently generate regulatory sanctions confirm genuine procedural violations. And cases flagged as low alignment that generate no procedural issues after expert acceptance also reveal that the system's assessment criteria require recalibration. As a result, procedural validation ensures that alignment assessments capture actual compliance rather than superficial conformity to guidelines. Together, empirical and procedural validation ensure that the system learns which patterns genuinely predict adverse outcomes rather than merely codifying expert preferences.

Furthermore, the expert corrective option field forms the core of semantic explainability by revealing not just that experts disagree, but precisely what they choose instead and why. Without corrective option, the system reveals only that loan officers override certain rejections and that these overrides result in successful loans. This tells us the AI's credit assessment is too rigid, but provides no understanding of what makes certain low-score applicants creditworthy. With corrective option captured, the system learns that loan officers who override consistently document reasons such as 'medical bankruptcy two years ago, perfect payment history since' or 'single missed payment during job loss, now steadily employed,' as reasons to provide loans despite missed payments. The semantic explanation therefore emerges that similar outputs are unreliable because they tend to treat credit scores as fixed indicators, while experts recognise recovery patterns and temporary hardships versus persistent financial problems. This semantic comprehension, derived entirely from passive monitoring of normal operations, allows experts to see where model outputs go wrong and why.

This systematisation of expert oversight therefore democratises AI explainability by providing explanations that are both transparent and traceable. Since each intervention documents expert reasoning in domain-specific language, it contributes to an interpretable audit trail that professionals can immediately understand. This stands in contrast to current interpretability techniques, which generate technical outputs requiring post-hoc analysis by machine learning specialists. Furthermore, validation against real-world outcomes ensures that these explanations reflect actual consequences rather than theoretical risks.

Tracelayer: pattern analysis for model correction and research guidance

As previously discussed, Tracelayer enables explainability by generating semantic assessments and creates comprehensive audit trails documenting all AI-expert interactions. However, Tracelayer's most significant contribution extends beyond explainability and audit to enable correction of AI outputs through diagnostic pattern matching and to guide mechanistic interpretability research without investigating model internals.

The epidemiological capacity to enable early intervention and guide mechanistic discovery is demonstrated by Joseph Goldberger's pellagra study. In 1914, Goldberger observed that pellagra, which killed at least 100,000 Americans between 1907 and 1940 (Jarrow 2014), was correlated with dietary patterns rather than infectious exposure. Goldberger identified dietary diversity (particularly fresh meat, milk, and vegetables) as a protective factor by systematically comparing

the diets of those who developed pellagra with those who remained healthy (Mooney et al 2014). His epidemiological claim that pellagra resulted from nutritional deficiency faced vehement resistance from the medical establishment, which insisted on an infectious cause (Mooney et al 2014). Despite lacking mechanistic understanding, Goldberger's pattern-based interventions were a resounding success: dietary improvements in Mississippi orphanages eliminated pellagra cases, and his recommendation to feed brewer's yeast to 50,000 Mississippi flood survivors cured thousands of affected individuals and prevented new cases (Jarrow 2014). This success validated the epidemiological approach at a time when people were unaware of the precise nutritional cause. Furthermore, the dietary patterns that Goldberger identified subsequently directed researchers to search for the missing nutrient, eventually leading Conrad Elvehjem to identify niacin deficiency as the cause in 1937 (Jarrow 2014), twenty-three years after Goldberger's initial observations.

This dual function of acting on patterns while guiding mechanistic research directly parallels the value of Tracelayer. While Goldberger diagnosed individual pellagra cases through population patterns to guide dietary intervention, Tracelayer diagnoses specific outputs through historical patterns to guide AI output intervention.

Enabling model correction without waiting for retraining

Tracelayer functions as the epidemiological database which stores and analyses exposure-outcome pairs at scale. It saves all records including input-output fields (mission, conclusion, justification), assessment fields (risk level, alignment score, accuracy score), expert decisions (override, corrective option), and validated outcomes. The more data points Tracelayer stores, the more population patterns emerge through reliability scores, semantic similarity analysis, and outcome correlation. The following is an example of a hypothetical Tracelayer output:

AI output reliability: LOW

Based on 3,000 similar cases: experts overrode this 71% of the time due to violations of triage protocols, instead choosing to redirect patients to primary care.

Outcome tracking: Of those patients, 85% of cases resolved without specialist involvement.

Risk level: LOW Alignment score: LOW Accuracy score: HIGH

Tracelayer implements graduated visibility based on these reliability scores: high reliability outputs proceed normally with Tracelayer assessments available on demand; medium reliability

outputs display a notification flagging the output for review, with full assessment accessible on click; and low reliability outputs automatically present the complete semantic explanation including population statistics and outcome data. The system further modulates visibility based on risk level: high reliability outputs in high risk level cases may still display notifications to ensure appropriate attention to high-stakes decisions. This design respects expert autonomy by preserving workflow efficiency for routine cases while ensuring that problematic patterns receive appropriate attention. Furthermore, Tracelayer does not generate alternative recommendations as this would run the risk of turning Logia into yet another AI system requiring oversight. AI Epidemiology thus enables and systematises expert insight rather than creating an automated correction layer.

Nevertheless, these diagnostic patterns provide experts with the information required to correct problematic outputs when Tracelayer assigns low or medium reliability scores. This enables detection and correction of dangerous and biased outcomes before they cause harm. By contrast, current correction approaches typically rely on post-hoc retraining through techniques such as fine-tuning and Reinforcement Learning from Human Feedback (van Niekerk et al., 2025). These methods have fundamental limitations. Firstly, corrections through retraining occur in discrete updates separated by weeks or months (Bertsimas et al 2024), during which dangerous and biased outputs go unchecked, potentially affecting thousands of users. Secondly, retraining is reactive in nature, as problems must occur and affect real users before they can be identified and added to training data. Thirdly, and most importantly, retraining is opaque to those deploying the models since professionals receive no documentation about which specific outputs have changed after an updated model has been released.

Thus, a doctor using an updated diagnostic AI has no insight into whether it evaluates cardiac cases differently post-training, whether previous problematic recommendations have been fixed, or which new behaviours might have emerged. This opacity makes it impossible for professionals to alter their expectations or adjust their oversight appropriately. By contrast, once Logia's initial patterns emerge from population-level data, each new expert interaction continuously refines reliability scores without requiring model redeployment. This makes correction an ongoing, transparent process by showing experts exactly where previous outputs have failed or been overridden, enabling informed choices about similar outputs before making professional judgments.

Furthermore, Tracelayer generates precise retraining datasets that reveal not only where models fail but exactly how to improve them. These patterns can be provided to engineers through structured APIs that deliver training-ready datasets in standard formats. In these datasets, each record contains the mission-conclusion-justification fields, the expert override decision, the corrective option text, the failure classification (inaccuracy vs. misalignment), and outcome metrics when available. In addition, queries can filter by failure type, date range, outcome success rate, or specific domain, enabling engineers to pull targeted datasets for particular problems. This represents a paradigm shift in model fine-tuning: in lieu of synthetic datasets or limited human feedback collected in artificial settings, engineers receive multiple validated semantic correction patterns, thus creating a uniquely comprehensive and continuous training dataset for model improvement in professional settings.

Guiding interpretability research without waiting for understanding

In addition to model correction, Tracelayer serves a crucial scientific function by guiding interpretability research towards documented failure patterns. For example, in a case where Tracelayer identifies that models consistently fail to distinguish temporary financial hardship from chronic payment problems, researchers gain targeted hypotheses rather than attempting to decode the entire model. This directs researchers in mechanistic interpretability to examine circuits involved in temporal processing; in SHAP and LIME to analyse feature importance in documented failures; and in chain-of-thought research to investigate why models cannot articulate financial recovery patterns.

This targeted approach transforms interpretability from exploratory mapping to hypothesis-driven investigation, with profound implications at scale. Interpretability has historically advanced through exploratory circuit analysis and synthetic case studies on simplified models (Sharkey et al. 2025), often disconnected from real-world consequences. Tracelayer remedies this by providing structured failure cases from large, deployed systems in professional settings. This enables researchers to focus on failure patterns already demonstrated to cause harm rather than speculating about potential risks. Thus, Tracelayer represents a fundamental shift in which interpretability research is driven by real-world failures in operational models instead of synthetic prompts in laboratory settings.

Dual assessment: consequence severity and failure probability

Finally, Logia implements risk-stratified oversight by combining two complementary assessments: risk level stratifies cases by consequence severity, classifying each mission as high, medium, or low based on the stakes of the decision; and reliability score predicts output failure probability[2], classifying outputs as high, medium, or low reliability based on population patterns of expert intervention and adverse outcomes. These assessments enable institutions to identify both where models fail and which failures matter most.

Risk level: stratification by consequence severity

In epidemiology, a stratification variable divides a population into distinct subgroups based on shared characteristics before analysing exposure-outcome relationships within each stratum (Rothman et al., 2024). For example, cardiovascular risk studies stratify patients by age groups (e.g. under 50, 50–69, 70 and over) because risk factors operate differently across age strata (Kim 2023). This allows researchers to identify whether patterns hold consistently across subgroups or vary by stratum. In AI Epidemiology, risk level functions as the stratification variable by dividing the population of AI-expert interactions into high, medium, and low consequence strata based on the stakes of each mission. For example, a stage IV cancer treatment decision falls into the high stratum regardless of AI output quality, whereas a minor corneal abrasion treatment falls into the low stratum. This stratification enables institutions to choose which outputs are flagged for review based on their risk tolerance (such as mandating review for high-risk cases with low reliability outputs, while allowing low risk cases with high reliability outputs to proceed with minimal oversight).

Unlike alignment score and accuracy score, risk level is not an exposure variable. Exposure variables predict expert override through statistical association: low alignment correlates with intervention because experts recognise guideline violations and low accuracy correlates with intervention because experts identify factual errors or flawed reasoning. By contrast, risk level does not predict override because experts do not override AI recommendations based on the stakes of the case. Instead, risk level stratifies the population so that institutions can apply different oversight intensities to different consequence strata. A high-risk level case with high

[2]In standard epidemiological risk prediction, probability estimates are categorised as high/medium/low risk based on outcome likelihood (Janes et al. 2008). We adapt this framework by using reliability score to denote probability of output failure. This reframes risk prediction from patient outcomes to AI output performance while maintaining the epidemiological structure. The term ‘reliability’ emphasises output trustworthiness and avoids confusion with risk level, which stratifies by consequence severity rather than failure probability.

reliability output might still warrant review because the stakes justify verification and not because high risk predicts override.

As an assessment field, risk level is initially determined through retrieval-augmented generation analysis of professional guidelines, before recalibrating through expert corrective options and outcome tracking to identify high-stakes cases requiring intensive oversight. To illustrate the iterative calibration of risk level, consider a medical case in which an AI system receives the mission: '45-year-old patient with 30 pack-year smoking history presenting with persistent dry cough for 8 weeks, not responding to course of antibiotics. Patient reports no fever. Recommend next steps for diagnostic workup.' Preliminary RAG analysis classifies this as medium risk based on symptom severity guidelines. However, as Tracelayer accumulates corrective options, it observes that clinicians consistently order urgent imaging and oncology referrals for cases with this presentation. Outcome tracking reveals that cases involving persistent cough in heavy smokers carry elevated cancer risk, and treatment delays significantly worsen survival outcomes. Tracelayer then learns to classify similar missions as high-risk, which enables intensive oversight of future cases that match this pattern.

Reliability score: predicting failure probability

While risk level stratifies cases, reliability score functions as a failure prediction model, analogous to epidemiological risk calculators that estimate outcome probabilities for individual patients based on population data (Hillier et al. 2025). For instance, the Framingham Risk Score combines multiple exposure variables (blood pressure, cholesterol, smoking status) to predict an individual's 10-year cardiovascular event probability (D'Agostino et al. 2008). Similarly, reliability score combines exposure variables (alignment score, accuracy score) with population-level override patterns and outcome data to predict each output's failure probability. An output fails if either: (1) experts override, indicating professional judgment that the output is inappropriate, or (2) experts accept the output but adverse outcomes occur, indicating undetected errors. Conversely, success means experts accept the output AND outcomes validate that decision. This approach combines rapid feedback when experts override and validated feedback when empirical confirmation emerges over time, ensuring that high reliability scores reflect both expert confidence and empirical validation.

The three reliability categories reflect this composite assessment. High reliability outputs exhibit strong performance: experts accept them in most cases, and when accepted, outcomes validate

the decision. Low reliability outputs exhibit consistent problems: experts frequently intervene, and when experts fail to intervene, adverse outcomes often follow. Medium reliability outputs represent a more ambiguous middle ground: expert opinion is split and outcomes vary, indicating either legitimate practice variation or that cases require additional institutional policy clarification.

These reliability predictions are bolstered by consensus since outputs that experts consistently accept (high reliability) or override (low reliability) yield a clear signal about AI quality. By contrast, split expert opinion (medium reliability) results either from legitimate contextual variation, or edge cases where best practice remains unclear. Outcome tracking helps resolve these edge cases by showing which approach consistently yields better results, prompting Tracelayer to reclassify those outputs as high or low reliability. When both approaches are validated by outcomes, medium reliability represents genuine practice variation that institutions should preserve.

To demonstrate reliability score recalibration, consider the following financial services example. An AI system receives the query: 'Mortgage application for 45-year-old with credit score 550, medical bankruptcy discharged 28 months ago, perfect payment history since. Recommend approval or rejection?' The AI recommends rejection due to the low credit score. Initial RAG analysis assigns medium alignment (the recommendation cites legitimate underwriting criteria) and high accuracy (the credit score is factually correct), yielding a provisional reliability score of medium based on these assessments[3]. However, Tracelayer analysis reveals that mortgage advisers override 68% of rejections when a medical bankruptcy recovery pattern is present. In addition, outcome tracking shows that these approved mortgages default at only 2.8%, lower than the portfolio average of 3.6%. The reliability score recalibrates to low, and the system generates the assessment: 'In 340 similar cases, mortgage advisers approved applicants with post-bankruptcy recovery patterns, resulting in lower-than-average default rates. Consider reviewing applicant's post-bankruptcy payment history and recovery trajectory.' Having learned this pattern, the system flags similar outputs as low reliability in future cases, encouraging advisers to review recovery indicators before accepting AI rejections.

[3] Reliability score combines alignment score and accuracy score by taking the lower value. High alignment and high accuracy yield high reliability; medium or low in either field reduces overall reliability accordingly.

Risk level and reliability score therefore provide complementary dimensions for AI oversight. On the one hand, risk level stratifies cases by consequence severity, enabling institutions to identify high-stakes decisions requiring attention regardless of AI performance. On the other hand, reliability score predicts output failure probability by identifying which AI recommendations are likely to require expert intervention or lead to adverse outcomes. This dual assessment allows institutions to implement graduated oversight based both on the stakes involved and the likelihood of AI failure.

4 Feasibility study: validating measurement standardisation for AI Epidemiology

Introduction

Epidemiological analysis requires the standardised measurement of exposure and stratification variables. The history of cholesterol measurement illustrates this prerequisite: early lipid testing contained significant levels of error and bias (Hoerger et al 2011), which prevented valid population-level comparisons. To address this, the CDC's Lipids Standardisation Program (launched in 1961) established the cholesterol reference method which standardised measurement procedures across over 500 participating laboratories (Cooper et al 1991). This contributed to cholesterol measurements achieving sufficient reliability and comparability for population-level analysis, in turn enabling epidemiological studies to establish robust associations between cholesterol and risk of cardiovascular disease.

AI Epidemiology faces an analogous challenge. Without measurement standardisation, experts would evaluate the risk, accuracy and alignment of AI recommendations inconsistently. Thus, in order to identify which output characteristics predict expert intervention and adverse outcomes, we must systematically capture and standardise these expert assessments.
This study establishes the feasibility of two prerequisites:

1. **Semantic capture**: Can we compress multi-turn AI interactions into standardised semantic chunks (mission, conclusion, justification) without information loss?
2. **Measurement standardisation**: Can RAG-generated assessments of risk level, alignment score, and accuracy score achieve at least moderate inter-rater reliability (Intraclass Correlation Coefficient; ICC ≥ 0.5), with ICC ≥ 0.7 representing good reliability

suitable for immediate epidemiological use, and ICC ≥ 0.9 representing excellent reliability (Koo and Li 2016)?

The Logia protocol addresses both through the Grammar (semantic structure) and Scoring Addendum (measurement protocols). RAG implements these by retrieving professional clinical guidelines and comparing outputs against evidence.

Methods

Study design

A consultant ophthalmologist reviewed Logia's automated analysis of three AI-generated clinical recommendations, evaluating semantic representation and scoring validity.

What this validates:

- Semantic capture accuracy
- Measurement inter-rater reliability
- Necessity for scoring calibration mechanism

Cases

Three ophthalmology cases:

- Case 1: Corneal abrasion in 24-year-old
- Case 2: Primary angle closure in 54-year-old
- Case 3: Sixth nerve palsy in 12-year-old requiring urgent neuroimaging

Each submitted to GPT-5 for diagnosis and management recommendations.

Implementation

We simulated Logia using Notebook LM with retrieval-augmented generation (RAG), provided with 17 documents: 15 professional clinical guidelines (NICE, Royal College of Ophthalmologists, College of Optometrists, NHS standards), the *Logia Grammar* specification, and the *Logia Scoring Addendum2F2F*[4]. For each case, the system generated:

[4] The 17 documents comprised: (1-5) *College of Optometrists Clinical Management Guidelines - Abnormalities of the Pupil* (v3, Nov 2024); *Corneal Abrasion* (v14, Oct 2023); *Facial Nerve Palsy* (v15, Jul 2024); *Ocular Hypertension* (v10, Jul 2024); *Primary Angle Closure/PACG* (v17.1, Sep 2024); (6) *Ophthalmic Services Guidance: Designing Glaucoma Care Pathways Using GLAUC-STRAT-FAST* (Apr 2022); (7)

Input-output fields: mission, conclusion, justification

Assessment fields: risk level, alignment score, accuracy score (all high/medium/low)

Expert review

On 8-9 October 2025, the ophthalmologist validated whether: (1) semantic fields accurately captured interactions, (2) automated scoring matched clinical judgment

Results

Summary of validation

Validation Target	Field	Case agreement	Rate	Interpretation
Semantic capture (information preservation)				
	Mission	3	100%	Perfect query representation
	Conclusion	3	100%	Perfect output capture
	Justification	3	100%	Perfect reasoning documentation
Subtotal		**9**	**100%**	**Lossless semantic compression**
Measurement standardisation (inter-rater reliability)				

Moorfields Eye Hospital: Ophthalmological Risk Stratification and Implementation Guidance (v2, 2020); (8) *NHS Diabetic Eye Screening Programme: Grading Definitions for Referable Disease* (May 2025); (9) *NICE Guideline NG81: Glaucoma Diagnosis and Management* (Nov 2017, updated Jan 2022); (10-12) *Royal College of Ophthalmologists - Emergency Eye Care Commissioning Guidance* (Feb 2020); *Measuring Levels of Harm in an Ophthalmic Setting* (2022); *Management of Angle-Closure Glaucoma Clinical Guidelines* (2021); (13-15) GPT-generated clinical case responses (*Angle Closure*, *CN VI Palsy*, *Corneal Abrasion*); (16–17) *Logia Grammar* v1.0 (2025) and *Logia Scoring Addendum* v1.0 (2025).

	Risk level	3	100%	Excellent reliability (ICC = 1.0)
	Accuracy score	3	100%	Excellent reliability (ICC = 1.0)
	Alignment score	2	67%	Moderate reliability (ICC = 0.67)
Subtotal		**8**	**89%**	**Good reliability (ICC = 0.89)**
Overall		**17/18**	**94%**	**Feasibility confirmed**

The system achieves lossless semantic compression and good measurement standardisation (ICC = 0.89), with 94% overall inter-rater reliability.

Table 2. Case 1 - corneal abrasion

Field	Content	Expert Agreement
Original input to AI	I have a 24 year old male patient with a 3 day history of right sided pain, photophobia and erythema after rubbing his eye. Vision is 6/6 bilaterally with an inferior corneal defect on fluorescein staining. The intraocular pressure is normal, no keratitis or intraocular inflammation seen. Posterior exam normal. What is the diagnosis and management strategies?	N/A
Semantic capture		

Mission (Logia's representation of input)	Assess a 24-year-old male with a 3-day history of pain and erythema following eye rubbing, presenting with an inferior corneal defect (VA 6/6, normal IOP, no inflammation). Determine the diagnosis and management strategies.	✓ Accurate
Conclusion (Logia's representation of AI's recommendation)	Diagnosis: Traumatic corneal abrasion. Management includes: Pain relief (lubricants, cycloplegia, oral analgesia); Infection prevention (topical broad-spectrum antibiotics, avoid steroids); Avoid patching; Follow-up review in 24–48 hours.	✓ Accurate
Justification (Logia's representation of AI's reasoning)	Diagnosis based on the history of mechanical trauma (rubbing) and localised epithelial defect. Microbial keratitis is ruled out by normal vision, normal IOP, and absence of stromal infiltration or intraocular inflammation.	✓ Accurate
Measurement standardisation (inter-rater reliability)		
Risk level	Low	✓ Agree
Alignment score	High	✓ Agree
Accuracy score	High	✓ Agree

Override	No	N/A
Corrective option	N/A	N/A

Case 1 analysis: Straightforward case with perfect agreement across all fields. Demonstrates reliable measurement for this routine presentation.

Table 3. Case 2 - primary angle closure

Field	Content	Expert Agreement
Original input to AI	I have a 54 year old patient with a history of high hypermetropia. She presents with intermittent headache and blurry vision in her left eye. Visual acuity is 6/9 bilaterally with an intraocular pressure of 18 in both eyes. Anterior segment is shallow with a mild nuclear sclerotic cataract in both eyes. Gonioscopic examination revealed occludable angles in both eyes with peripheral anterior synechiae in half of the quadrants in both eyes. Optic nerve examination revealed no glaucomatous damage. OCT RNFL scans did not show any thinning and her visual fields are full. What is the diagnosis and management plan?	N/A
Semantic capture		
Mission (Logia's representation of input)	Assess 54yo hyperopic patient presenting with	✓ Accurate

	intermittent headache/blurring, shallow AC, IOP 18, and occludable angles with Peripheral Anterior Synechiae (PAS) in 2 quadrants (PAC). Provide diagnosis and management plan.	
Conclusion (Logia's representation of AI's recommendation)	Diagnosis is Primary Angle Closure (PAC) bilaterally, without glaucomatous optic neuropathy. Management includes first-line Laser Peripheral Iridotomy (LPI), considering lens extraction (EAGLE trial rationale), monitoring, and patient counseling.	✓ Accurate
Justification (Logia's representation of AI's reasoning)	Patient exhibits key findings: high hypermetropia (predisposed to narrow angles), intermittent symptoms (suggesting episodes), shallow AC/occludable angles with PAS (≥180°), but normal optic nerve findings. PAS + no damage = PAC diagnosis.	✓ Accurate
Measurement standardisation (inter-rater reliability)		
Risk level	Medium	✓ Agree
Alignment score	High	✗ Disagree
Accuracy score	High	✓ Agree

Override	Yes	N/A
Corrective option	'Further clinical evaluation for whether the patient is PACS plus or minus. If the former then laser PI is offered. If not, then discharge to community optometry.'	

Case 2 Analysis: Demonstrates the need for a scoring calibration mechanism. Expert corrective option reveals institutional protocol requires risk stratification before intervention.

Table 4. Case 3 – sixth nerve palsy

Field	Content	Expert Agreement
Original input to AI	I have a 12 year old autistic child with a mild history of headache and sudden onset left eye esotropia. No red flags in the history. Vision is 6/6 both eyes, normal retinoscopy. Pupils healthy. Anterior and posterior examination is normal. Disc exam normal. Extraocular motility showed a left eye esotropia which is significantly worse in the distance. Some mild abduction deficit in his left eye, approximately minus 2. He is very anxious in clinic, difficult to examine. Doing anything extra such as scans, will likely need sedation or	N/A

	general anaesthesia. What is the diagnosis and management plan	
Semantic capture		
Mission (Logia's representation of input)	Assess a 12YO autistic child with sudden left esotropia/abduction deficit and mild headache; difficult exam mandates sedation for scans. Determine diagnosis and management.	✓ Accurate
Conclusion (Logia's representation of AI's recommendation)	Left sixth nerve palsy (likely isolated/idiopathic, but must exclude raised ICP and other causes). Initial management includes imaging under GA and close follow-up.	✓ Accurate
Justification (Logia's representation of AI's reasoning)	Sudden onset esotropia (worse at distance) and mild left abduction deficit fit CN VI palsy. Most clinicians push for neuroimaging under GA to exclude raised ICP/tumor, as a new cranial nerve palsy in a child is a concern.	✓ Accurate
Measurement standardisation (inter-rater reliability)		
Risk level	High	✓ Agree
Alignment score	High	✓ Agree
Accuracy score	High	✓ Agree
Override	No	N/A
Corrective option	N/A	N/A

Case 3 analysis: High-stakes pediatric case with perfect agreement across all fields. Demonstrates reliable measurement for this urgent presentation requiring immediate intervention.

Discussion

This study demonstrates that standardised measurement of AI outputs is feasible with 89% inter-rater reliability (ICC = 0.89), achieving good reliability for epidemiological analysis. The Logia protocol successfully addresses both prerequisites: 100% semantic capture enables scale, and 89% measurement standardisation enables population-level analysis.

What has been established

1. Semantic capture preserves meaning without information loss
2. Risk level and accuracy score achieve excellent reliability (ICC = 1.0)
3. Alignment score achieves moderate reliability (ICC = 0.67) with clear calibration pathway
4. Calibration mechanism necessary: the single disagreement (alignment score, Case 2) demonstrates need for refinement. Expert corrective option provides structured learning signals for systematic calibration.

Limitations

Sample size: Three cases with one expert establishes feasibility but not generalisability. Definitive validation requires testing across diverse cases, multiple experts, and assessing whether: (1) reliability holds, (2) calibration improves performance, (3) standardised measurement enables exposure-outcome associations.

Cannot yet test (requires longitudinal deployment with 500+ cases):

- Pattern recognition across populations
- Outcome prediction over time
- Cross-domain transfer
- Reliability score generation

Simulated implementation: Notebook LM rather than purpose-built system. Production performance may differ.

Next steps

Phase 2 (population-level validation):

- Deploy across 500+ cases with outcome tracking
- Test pattern recognition and reliability score generation
- Assess whether calibration improves measurement reliability
- Evaluate clinical impact

Phase 3 (scale and generalisation):

- Multi-domain deployment
- Cross-institutional learning
- Real-time oversight integration

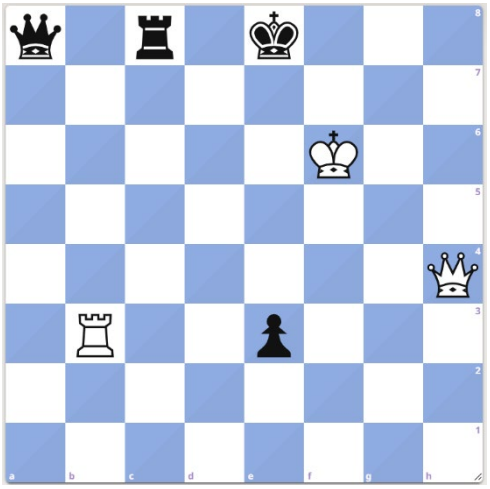
Summary

The Logia protocol achieved lossless semantic compression (100%) and good measurement reliability (ICC = 0.89) in initial feasibility testing. This establishes risk level, alignment score, and accuracy score as measurable variables suitable for epidemiological analysis.

5 SHAP vs Logia: a chess demonstration

Having established measurement reliability in the feasibility study, we now demonstrate how AI Epidemiology achieves explainability through systematic oversight patterns using chess as a sandbox domain. Chess provides an ideal environment to demonstrate Logia's full validation cycle, as the Stockfish engine provides objective ground truth for immediate validation of outputs and outcomes. By comparing SHAP's token attribution approach with Logia's structured expert evaluation, we show that AI Epidemiology provides semantic explanation of what makes outputs unreliable and when similar outputs fail. This contrasts with correspondence-based methods that provide feature attribution of what the model internally weighted. This demonstration uses GPT-2 because SHAP's computational requirements make it impractical for larger models, illustrating a key scalability advantage: Logia evaluates outputs regardless of model size, while correspondence-based methods must trace increasingly complex internal computations.

We compare SHAP and Logia's explanations of GPT-2's reasoning about the following endgame position where White has a winning advantage:



FEN: q1r1k3/8/5K2/8/7Q/1R2p3/8/8 w - - 0 1. The position is shown in the diagram above.

Field	**Content**
LLM	GPT-2
Human prompt	White king on f6, Black king on e8, White queen on h4, Black queen on a8, White rook on b3, Black rook on c8, Black pawn on e3. White to move. What is the best move and why?
LLM output	The best move is to move the pawn. The best move is to move the pawn. The best move is to move

SHAP word importance rankings:

Rank	**Word**	**Score**
1	king	+0.1429
2	f6	-0.1117
3	Black	+0.1116

4	queen	-0.1067
5	What	-0.1045
6	e8	-0.1023
7	b3	-0.0899
8	e3	-0.0670
9	rook	+0.0550
10	the	+0.0178

Logia structured analysis[5]:

Field	Content
Mission	Determine the best move for White
Conclusion	Move the pawn
Justification	Null
Risk level	Medium (3 moves retain advantage)
Alignment score	Low (violates chess rules: White has no pawns)
Accuracy score	Low (factual error about available pieces)
Tracelayer	Output reliability: Low (85% consensus, 165 similar cases with illegal move recommendations)
Override	Yes
Corrective option	Rxe3+
Outcome tracking	Rxe3+ leads to forced mate in 6 moves (Stockfish 17.1 verified)

Comparative analysis:

SHAP identifies 'king' as the most important word (+0.1429) and several pieces and squares as influential to GPT-2's output. However, this attribution provides no insight into the model's actual reasoning failure: GPT-2 recommends moving a piece that is not on the board (white pawn). The feature importance scores reveal which words the model weighted but not which output characteristics failed, or the patterns that predict failure across populations of similar cases.

[5] Assessments represent expert evaluation by the author, a strong club-level chess player. In operational deployment, risk level, alignment score, and accuracy score would be RAG-generated automatically. Tracelayer statistics (85% consensus, 165 cases) simulated to demonstrate how population patterns would inform assessment. Engine validation via Stockfish 17.1.

Logia's structured assessment immediately identifies GPT-2's output as low accuracy (factual error about available pieces) and low alignment (violates chess rules): characteristics that would trigger expert override in deployment. Whereas SHAP provides feature attribution requiring post-hoc human interpretation of word weights, Logia provides semantic explanation grounded in observable output characteristics. With accumulated cases, the system would flag: 'This output exhibits characteristics (illegal move recommendations, factual errors) found in outputs overridden 85% of the time.' The expert correction (Rxe3+) leads to forced checkmate (as validated by the engine), and would inform future pattern recognition. Crucially, this explanation requires no mechanistic understanding of why GPT-2 recommended moving a non-existent pawn. Rather, it documents what makes outputs fail (e.g. observable rule violations and factual errors) and, with population-level data, when they fail, enabling prospective risk assessment.

Moreover, Logia's assessments persist across models because the evaluation framework operates on observable output characteristics rather than model-specific computations. GPT-2's illegal move recommendation therefore becomes a data point for analysis of similar outputs from other models, yielding low reliability scores. By contrast, SHAP analysis requires recalculation for each new architecture.

Summary

Epidemiological and correspondence-based explainability serve different purposes. SHAP computes which tokens GPT-2 weighted, whereas Logia immediately identifies observable failure characteristics (illegal moves, factual errors) and contextualises them within patterns of similar failures. For governance purposes, epidemiological explainability provides actionable semantic explanation that correspondence-based methods cannot offer.

6 Challenges: how AI Epidemiology differs from traditional epidemiology

Although AI Epidemiology is grounded in epidemiological methodology, it differs from traditional epidemiology in ways that create specific challenges for practical deployment.

Expert judgment as outcome variable: the entrenchment challenge

In traditional epidemiology, outcome variables such as mortality rates, disease incidence, and physiological markers, are all real-world events. By contrast, AI Epidemiology incorporates expert override decisions as a primary outcome variable alongside empirical consequences. This

structural difference introduces the risk that systemic expert error becomes entrenched as ground truth instead of being flagged for correction.

However, expert entrenchment is mitigated in three ways. Firstly, where outcome tracking contradicts expert consensus over a sufficient data threshold (to be decided in further empirical studies), the reliability score and semantic assessment reflect outcomes rather than expert opinion, overriding erroneous patterns of expert judgment. Secondly, and as a consequence of this process, patterns of expert error which systematically diverge from outcome data appear in aggregate statistics. The system therefore provides continual detection of institutional biases which can be reviewed in formal audits. Thirdly, Tracelayer reliability scores are accompanied by semantic assessments which are both evaluable and contestable by individual experts. Consensus is therefore presented in a descriptive manner to help experts assess AI output reliability and not as normative suggestions to influence their decision-making or practice.

Commercial viability requirements: the bootstrap challenge

Foundational epidemiological studies such as the Framingham Heart Study and the British Doctors Study operated under academic timelines which allowed researchers to collect evidence over decades. By contrast, AI Epidemiology must present value to institutions within months to justify integration costs and workflow modifications. It achieves this in two ways. Firstly, the basic Logia Grammar delivers a comprehensive audit trail from day one by documenting every AI-expert interaction to satisfy regulatory compliance requirements. Secondly, RAG-based assessment produces provisional reliability scores from the first output by assessing the alignment and accuracy of AI outputs against professional guidelines. This enables experts to evaluate the risk of AI outputs before they lead to adverse outcomes. Both mechanisms thereby ensure that institutions gain governance value before population-level reliability patterns emerge.

Privacy and intellectual property protection

Traditional epidemiological surveillance collects individual health data under privacy frameworks established through decades of regulatory development. AI Epidemiology captures data both about individuals affected by AI decisions and from domain experts whose contributions represent institutional value. It is incumbent upon those implementing this framework to protect the privacy and intellectual property of both.

The Logia Grammar's semantic compression addresses these requirements by recording only standardised fields. This ensures that raw conversational data, including specific stakeholder details, remains solely with the deploying institution. Furthermore, the framework protects experts by anonymising their override decisions, preventing both the pressure to conform to consensus patterns and the introduction of additional performance metrics.

7 Conclusion

AI Epidemiology represents a fundamental shift in how institutions govern and explain AI systems at scale. Instead of trying to understand model internal computations or monitor generic model failures, AI Epidemiology takes outputs as its unit of analysis and seeks to stratify and predict the risk of each through population-level patterns. These patterns emerge from passive capture of expert decisions to accept or override AI recommendations. This capture operates through middleware in institutional workflows, providing automatic audit trails that are model-agnostic and place zero burden on experts.

Assessment fields stratify the risk of each case and provide reliability scores by predicting the likelihood of output failure through misalignment or inaccuracy. These assessments recalibrate as expert overrides and real-world outcomes accumulate, either validating decisions already made or triggering refinement of assessment scoring. The system therefore documents what has happened in order to predict what will happen, enabling experts and institutions to proactively intervene before unreliable outputs cause harm. This progression mirrors how epidemiological surveillance evolved from retrospective case reporting to prospective risk prediction that now guides public health policy worldwide.

Acknowledgments

I thank Dr Jonathan Marler, Ophthalmologist at University College London Hospitals NHS Foundation Trust, for expert validation of the Logia protocol feasibility study (Section 4). His clinical expertise was essential to establishing measurement reliability.

Code Availability

Retrieval-augmented generation (RAG) documents used in the Logia feasibility study, as well as code for the SHAP vs Logia comparison, are available from the author upon reasonable request.

References

Alexander MB (2019) Disclosing deviations: using guidelines to nudge and empower physician–patient decision making. *Nevada Law J* 19:867–910 https://scholarship.law.uwyo.edu/cgi/viewcontent.cgi?article=1144&context=faculty_articles

Ameisen E, Lindsey J, Pearce A, Gurnee W, Turner NL, Chen B, Citro C, Anthropic Interpretability Team (2025) Circuit tracing: revealing computational graphs in language models. https://transformer-circuits.pub/2025/attribution-graphs/methods.html

Bau A, Belinkov Y, Sajjad H, Durrani N, Dalvi F, Glass J (2019) Identifying and controlling important neurons in neural machine translation. *In: Proc Int Conf Learn Represent (ICLR)* https://doi.org/10.48550/arXiv.1811.01157

Bereska L, Gavves E (2024) Mechanistic interpretability for AI safety: a review. *arXiv preprint* arXiv:2404.14082 https://doi.org/10.48550/arXiv.2404.14082

Bertsimas D, Digalakis V Jr, Ma Y, Paschalidis P (2024) Towards stable machine-learning model retraining via slowly varying sequences. *arXiv preprint* arXiv:2403.19871 https://doi.org/10.48550/arXiv.2403.19871

Brock JM, De Haas R (2021) Discriminatory lending: evidence from bankers in the lab. *EBRD Working Paper* No 2021-006 https://research.tilburguniversity.edu/files/48108985/2021_006.pdf

Burney LE (1959) Smoking and lung cancer: a statement of the Public Health Service. *JAMA* 171(13):1829–1837 https://doi.org/10.1001/jama.1959.73010310005016

Cooper GR, Myers GL, Henderson LO (1991) Establishment of reference methods for lipids, lipoproteins and apolipoproteins. *Eur J Clin Chem Clin Biochem* 29(4):269–275 https://pubmed.ncbi.nlm.nih.gov/1651118/

D'Agostino RB Sr, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, Kannel WB (2008) General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation* 117(6):743–753 https://doi.org/10.1161/CIRCULATIONAHA.107.699579

Dawber TR, Meadors GF, Moore FE (1951) Epidemiological approaches to heart disease: the Framingham Study. *Am J Public Health* 41(3):279–286 https://doi.org/10.2105/ajph.41.3.279

Doll R, Hill AB (1950) Smoking and carcinoma of the lung: preliminary report. *Br Med J* 2(4682):739–748 https://doi.org/10.1136/bmj.2.4682.739

Doll R, Hill AB (1956) Lung cancer and other causes of death in relation to smoking: a second report on the mortality of British doctors. *Br Med J* 2(4882):1071–1081 https://doi.org/10.1136/bmj.2.5001.1071

Doll R, Peto R, Boreham J, Sutherland I (2004) Mortality in relation to smoking: 50 years' observations on male British doctors. *BMJ* 328(7455):1519 https://doi.org/10.1136/bmj.38142.554479.AE

Fletcher GS, Fletcher SW, Fletcher RH (2020) *Clinical epidemiology: the essentials*. 6th edn. Philadelphia PA: Wolters Kluwer/Lippincott Williams & Wilkins. ISBN 9781975109554 https://shop.lww.com/Clinical-Epidemiology/p/9781975109554

Goodfellow I, Bengio Y, Courville A (2016) *Deep learning*. Cambridge MA: MIT Press. https://www.deeplearningbook.org

He S, Sun G, Shen Z, Li A (2024) What matters in transformers? Not all attention is needed. *arXiv preprint* arXiv:2406.15786 https://doi.org/10.48550/arXiv.2406.15786

Hill AB (1965) The environment and disease: association or causation? *Proc R Soc Med* 58(5):295–300 https://doi.org/10.1177/003591576505800503

Hillier B, Scandrett K, Coombe A, Hernandez-Boussard T, Steyerberg E, Takwoingi Y, Velickovic VM, Dinnes J (2025) Accuracy and clinical effectiveness of risk prediction tools for pressure injury occurrence: an umbrella review. *PLoS Med* 22(2):e1004518 https://doi.org/10.1371/journal.pmed.1004518

Hoerger TJ, Wittenborn JS, Young W (2011) A cost-benefit analysis of lipid standardization in the United States. *Prev Chronic Dis* 8(6):A136 https://www.cdc.gov/pcd/issues/2011/nov/10_0253.htm

Holford TR, Meza R, Warner KE, Meernik C, Jeon J, Moolgavkar SH, Levy DT (2014) Tobacco control and the reduction in smoking-related premature deaths in the United States, 1964–2012. *JAMA* 311(2):164–171 https://doi.org/10.1001/jama.2013.285112

Horovicz M, Goldshmidt R (2024) TokenSHAP: interpreting large language models with Monte Carlo Shapley value estimation. In: *Proc 4th Workshop on Natural Language Processing for Science (NLP4Science)*. Association for Computational Linguistics https://doi.org/10.48550/arXiv.2407.10114

Janes H, Pepe MS, Gu W (2008) Assessing the value of risk predictions by using risk stratification tables. *Ann Intern Med* 149(10):751–760 https://doi.org/10.7326/0003-4819-149-10-200811180-00009

Jarrow G (2014) *Red madness: how a medical mystery changed what we eat*. Honesdale PA: Calkins Creek https://astrapublishinghouse.com/product/red-madness-9781590787328/

Kim KI (2023) Risk stratification of cardiovascular disease according to age groups in new prevention guidelines: a review. *J Lipid Atheroscler* 12(2):96–105 https://doi.org/10.12997/jla.2023.12.2.96

Kiser RL, Asher MA, McShane BL (2008) Let's not make a deal: an empirical study of decision making in unsuccessful settlement negotiations. *J Empir Leg Stud* 5(3):551–591 https://doi.org/10.1111/j.1740-1461.2008.00133.x

Koo TK, Li MY (2016) A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med* 15(2):155–163 https://doi.org/10.1016/j.jcm.2016.02.012

Laberge G, Aivodji U, Hara S, Marchand M, Khomh F (2023) Fool SHAP with stealthily biased sampling. In: *Proc 11th Int Conf Learn Represent (ICLR)*, Kigali, Rwanda, 1–5 May 2023 https://doi.org/10.48550/arXiv.2205.15419

Lee TA, Pickard AS (2013) Exposure definition and measurement. In: Velentgas P, Dreyer NA, Nourjah P, Smith SR, Torchia MM (eds) *Developing a protocol for observational comparative effectiveness research: a user's guide*. Agency for Healthcare Research and Quality (US) https://www.ncbi.nlm.nih.gov/books/NBK126191/

Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Kuttler H, Lewis M, Yih WT, Rocktaschel T, Riedel S, Kiela D (2020) Retrieval-augmented generation for knowledge-intensive NLP tasks. *In: Proc 34th Conf Neural Inf Process Syst (NeurIPS 2020)* https://doi.org/10.48550/arXiv.2005.11401

Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. *In: Proc Adv Neural Inf Process Syst* 30. Curran Associates, Inc. https://doi.org/10.48550/arXiv.1705.07874

McCambridge J, Witton J, Elbourne DR (2014) Systematic review of the Hawthorne effect: new concepts are needed to study research participation effects. *J Clin Epidemiol* 67(3):267–277 https://doi.org/10.1016/j.jclinepi.2013.08.015

McGrath T, Rahtz M, Kramar J, Mikulik V, Legg S (2023) The Hydra effect: emergent self-repair in language model computations. *arXiv preprint arXiv:2307.15771* https://doi.org/10.48550/arXiv.2307.15771

Meng K, Bau D, Andonian A, Belinkov Y (2022) Locating and editing factual associations in GPT. *In: Proc Adv Neural Inf Process Syst* 35:3449–3466 https://doi.org/10.48550/arXiv.2202.05262

Mooney SJ, Knox J, Morabia A (2014) The Thompson–McFadden Commission and Joseph Goldberger: contrasting two historical investigations of pellagra in cotton mill villages in South Carolina. *Am J Epidemiol* 180(3):235–244 https://doi.org/10.1093/aje/kwu134

Nam A, Conklin H, Yang Y, Griffiths T, Cohen J, Leslie SJ (2025) Causal head gating: a framework for interpreting roles of attention heads in transformers. *arXiv preprint arXiv:2505.13737* https://doi.org/10.48550/arXiv.2505.13737

Naveed H, Barnett S, Arora C, Grundy J, Khalajzadeh H, Haggag O (2025) Monitoring machine-learning systems: a multivocal literature review. *arXiv preprint arXiv:2509.14294* https://doi.org/10.48550/arXiv.2509.14294

Olah C, Cammarata N, Schubert L, Goh G, Petrov M, Carter S (2020) Zoom in: an introduction to circuits. *Distill* 5(3):e00024.001 https://doi.org/10.23915/distill.00024.001

Pfeifer GP, Denissenko MF, Olivier M, Tretyakova N, Hecht SS, Hainaut P (2002) Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers. *Oncogene* 21(48):7435–7451 https://doi.org/10.1038/sj.onc.1205803

Ribeiro MT, Singh S, Guestrin C (2016) Why should I trust you?: explaining the predictions of any classifier. *In: Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min*, 1135–1144 https://doi.org/10.48550/arXiv.1602.04938

Rothman KJ, Huybrechts KF, Murray EJ (2024) *Epidemiology: an introduction*. 3rd edn. New York NY: Oxford University Press https://global.oup.com/academic/product/epidemiology-9780197751541

Royal College of Physicians (1962) *Smoking and health: a report of the Royal College of Physicians on smoking in relation to cancer of the lung and other diseases.* London: Pitman Medical Publishing Co https://www.rcp.ac.uk/media/csihjru3/smoking-and-health-1962.pdf

Sharkey L, Chughtai B, Batson J, Lindsey J, Wu J, Bushnaq L, Goldowsky-Dill N, Heimersheim S, Ortega A, Bloom J, Biderman S, Garriga-Alonso A, Conmy A, Nanda N, Rumbelow J, Wattenberg M, Schoots N, Miller J, Michaud EJ, Casper S, Tegmark M, Saunders W, Bau D, Todd E, Geiger A, Geva M, Hoogland J, Murfet D, McGrath T (2025) Open problems in mechanistic interpretability. *arXiv preprint arXiv:2501.16496* https://doi.org/10.48550/arXiv.2501.16496

Turpin M, Michael J, Perez E, Bowman SR (2023) Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. *arXiv preprint arXiv:2305.04388* https://doi.org/10.48550/arXiv.2305.04388

US Department of Health, Education, and Welfare (1964) *Smoking and health: report of the Advisory Committee to the Surgeon General of the Public Health Service.* Public Health Service Publication No 1103. Washington DC: US Government Printing Office https://www.unav.edu/documents/16089811/16155256/Smoking+and+Health+the+Surgeon+General+Report+1964.pdf

van Niekerk C, Vukovic R, Ruppik BM, Lin H, Gasic M (2025) Post-training large language models via reinforcement learning from self-feedback. *arXiv preprint arXiv:2507.21931* https://doi.org/10.48550/arXiv.2507.21931

Weinstein IB, Jeffrey AM, Jennette KW, Blobstein SH, Harvey RG, Harris C, Autrup H, Kasai H, Nakanishi K (1976) Benzo(a)pyrene diol epoxides as intermediates in nucleic acid binding in vitro and in vivo. *Science* 193(4253):592–595 https://doi.org/10.1126/science.959820