

# **Towards AI epidemiology: A measurement standardisation framework for prospective risk detection**

## **Abstract**

This paper proposes AI Epidemiology as a measurement standardisation framework that compresses expert-AI interactions into standardised, comparable fields for model-free prospective risk detection in deployed AI systems. The main aim of this concept paper is to define the scope of the framework both semantically and statistically, so that empirical testing will be possible in future research. The population-level reliability claims that the framework is designed to support are therefore the subject of a staged research programme rather than results the paper claims to have established. The framework advances three claims in sequence: first, that LLMs under bounded conditions can assess evidential and policy alignment of expert-AI interactions; second, that these LLM-judged alignment scores can predict expert override as a signal of output failure; and third, that conditional on (1) and (2), alignment scores can act as exposure variables which predict downstream outcomes, reflecting epidemiological risk detection. This paper addresses the first claim and specifies the protocol for the second and third.

We acknowledge the structural circularity in using a black-box model to judge the alignment of other black-box model outputs and propose a partial escape via bounded conditions including structured inputs, explicit rubrics and reference documents. These reduce stochastic variance and noise. Systematic biases such as sycophancy, position bias, and self-preference are addressed through a reliability verification procedure described in Section 4.

The paper proposes the grammar, comprising three input-output fields (mission, conclusion, justification), one stratification variable (risk level), and two alignment scores (evidential alignment, policy alignment), each with field extraction and scoring rubrics. In addition, it comprises two governance fields (override and corrective option) captured from expert behaviour within the interaction. The evaluation protocol specifies DeLong's test, paired bootstrap inference, a pre-specified non-inferiority margin  $\delta = 0.05$ , and Holm-Bonferroni correction.

## **1 Introduction**

Modern AI systems present a fundamental governance challenge (Anderljung et al. 2023; UN High-Level Advisory Body on AI 2024). 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 (Burrell 2016; Bommasani et al. 2021). This opacity becomes critical when such models produce outputs that are dangerous, unethical, and inaccurate (Doshi-Velez and Kim 2017). There is

therefore an urgent need to detect AI output failures in a way that is both model-free and scalable.

AI Epidemiology is the systematic application of population-level surveillance methods to expert-AI interactions, in which structured measurements of AI outputs are aggregated across cases to identify patterns of evidential and policy misalignment. These patterns are then validated against expert overrides and, ultimately, downstream outcomes.

We introduce the term 'correspondence-based interpretability' to refer to methods which seek to establish correspondence between what the model focuses on internally and what it outputs. This includes mechanistic interpretability, feature attribution methods like SHAP (Lundberg and Lee 2017) and LIME (Ribeiro et al. 2016), and chain-of-thought prompting which presupposes that stated reasoning reflects actual computation (Turpin et al. 2023). These methods become computationally intractable and epistemically unreliable as models scale: attributions become unstable (Bereska and Gavves 2024; Mueller et al. 2025), and models generate plausible-sounding explanations that do not reflect their actual decision processes. While establishing 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 the governance gap described above, this paper proposes AI Epidemiology as a framework that bypasses correspondence-based interpretability by applying population-level surveillance methods from public health, identifying AI risks through systematic observation of outputs and expert interventions without requiring transparency into model computations. The framework asks which outputs are evidentially weak or misaligned with institutional policy, which observable characteristics predict these patterns, and where failure modes concentrate across domains, jurisdictions, and models.

The framework is scoped to regulated institutional settings in which AI outputs inform governance-relevant decisions including medicine, finance, legal practice, and public-sector decision-making. The scope is restricted to applications where prospective risk detection is required, recognising that in many machine-learning settings predictive performance against held-out data is a sufficient measure. The framework applies specifically to settings meeting three jointly necessary conditions, as specified in Section 10: an identifiable evidence base, an applicable policy corpus, and accountable human reviewers with observable override behaviour.

The paper develops four elements across the sections that follow. First, the grammar: a structured capture protocol with field extraction protocols and scoring rubrics. Second, the LLM-as-judge foundation: an argument that the grammar satisfies the operating conditions for reliable automated scoring, with bounded conditions addressing stochastic variance and a reliability verification procedure detecting residual systematic biases. Third, the

statistical specification: an evaluation protocol with DeLong's test and  $\delta = 0.03$  pre-specified. Fourth, the empirical programme: three stages through which the framework progresses.

These elements are developed in sequence. Section 2 situates AI Epidemiology within the broader context of interpretability research and adjacent literatures. Section 3 operationalises the grammar with extraction and scoring rubrics. Section 4 presents the LLM-as-judge foundation. Section 5 presents the three-claim argument structure and addresses the structural circularity. Section 6 specifies the evaluation protocol. Section 7 specifies iterative grammar refinement. Section 8 examines threats to validity. Section 9 describes the three-stage future programme. Section 10 specifies the domain of applicability. Section 11 concludes.

## **2 Beyond correspondence: situating AI epidemiology**

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. This enables a model-free form of AI output risk detection. Mechanistic interpretability research pursues the worthy goal of understanding how neural networks compute, but as public health does not rely solely on biological understanding, so too can AI risk detection proceed without complete mechanistic understanding.

### **The Bradford Hill precedent**

Austin Bradford Hill and Richard Doll demonstrated that statistical evidence could justify public health action without mechanistic understanding. Their 1950 case-control study (Doll and Hill 1950) and the British Doctors Study (Doll and Hill 1956; Doll et al. 2004) established a nine- to thirty-fold increased risk of lung cancer among smokers with clear dose-response relationships (Hill 1965). Public health action followed in the form of the 1964 US Surgeon General's report (U.S. Department of Health, Education, and Welfare, 1964) and subsequent regulation, while the molecular mechanisms by which tobacco smoke causes cancer remained unknown for decades thereafter (Weinstein et al. 1976). This methodological principle translates to AI Epidemiology, which assesses statistical risk of AI outputs being misaligned or evidentially weak.

However, epidemiological analysis of this kind requires the standardised measurement of AI outputs before population-level associations can be established. In classic epidemiology, early lipid testing contained significant error and bias (Hoerger et al. 2011), which prevented valid population-level comparisons until the CDC's Lipids Standardisation Program established a common reference method across over 500 laboratories (Cooper et al. 1991). Only once measurement was standardised could

associations between cholesterol and cardiovascular risk be reliably established. Similarly, population-level risk detection of AI outputs will only be possible through standardised measurement of their evidential and policy alignment. The grammar addresses this by compressing interactions into structured, comparable fields.

### **From correspondence to risk stratification: an epistemic shift**

The correspondence model seeks to make AI models transparent by mapping out a causal pathway from input to output through internal model computations. SHAP tests different combinations of words to identify feature contributions, but it is computationally impossible to test all possible combinations in an LLM containing millions of tokens, forcing approximations which can be unstable and misleading (Horovitz and Goldshmidt 2024). Furthermore, adversaries can manipulate SHAP explanations without altering model outputs: one attack reduced the apparent importance of a race input by 90%, allowing a biased model to pass a fairness audit (Laberge et al. 2023). In transformer models, both observational and interventional mechanistic interpretability methods face fundamental challenges: superposition and polysemanticity lead to individual components encoding multiple unrelated concepts (Olah et al. 2020; Bereska and Gavves 2024), and functions are distributed redundantly across components (He et al. 2024; McGrath et al. 2023), making causal attribution unreliable. Comprehensive validation through introspective faithfulness therefore remains a long-term research goal rather than a practical solution for urgent governance needs.

AI Epidemiology asks different questions, including: which outputs are misaligned with institutional policy? Which observable characteristics predict this misalignment, and where do alignment failures concentrate across domains, jurisdictions, and models? Observable characteristics which predict these patterns include weak evidential grounding and low policy alignment, such as an output recommending against institutional guidelines without citing appropriate evidence. This would score low on both criteria and predict expert override.

AI Epidemiology therefore shifts the locus of risk detection from internal processes to observable patterns and their consequences. For example, it might convey that an AI oncology recommendation contradicts treatment guidelines and resembles a pattern which was overridden 75% of the time by experts. Both approaches serve legitimate governance purposes on different timescales. Whereas correspondence-based interpretability pursues the long-term scientific goal of model transparency; AI Epidemiology pursues the governance goal of identifying risk and enabling appropriate oversight at scale.

### **AI Epidemiology vs adjacent literatures**

AI Epidemiology can be situated against several adjacent literatures, each of which it complements rather than displacing.

The relationship between AI Epidemiology and statistical monitoring tools mirrors that between epidemiology and disease surveillance. Statistical model monitoring tools such as IBM Watson OpenScale track model performance metrics analogously to disease surveillance tracking aggregate incidence rates (Naveed et al. 2025). By contrast, AI Epidemiology applies population-level misalignment patterns to individual outputs, analogously to clinical epidemiology using population patterns to inform decisions about individual patients (Fletcher et al. 2020). The two modalities are complementary: aggregate monitoring detects overall drift, while AI Epidemiology enables prospective triage at the output level.

Human-in-the-loop research designs systems in which human feedback shapes model behaviour through active learning, interactive machine learning, and machine teaching, so that human judgment is incorporated into the model's outputs over time (Mosqueira-Rey et al. 2023). Where HITL uses human feedback as a training signal, AI Epidemiology instead uses the record of human override behaviour as a validation signal that tracks expert recognition of misalignment. AI Epidemiology is therefore a risk detection layer which sits on top of the model rather than a training initiative which changes it.

Algorithmic auditing establishes procedures for examining AI systems against specified standards. For example, Raji et al. (2020) propose SMACTR: an end-to-end internal audit framework which generates documentation at each stage of the AI development lifecycle to support accountability. Such audits are typically post-hoc since they examine what a system has produced during periodic reviews. AI Epidemiology differs by accumulating structured interaction records for risk detection at deployment. This addresses Mökander et al.'s concern about the absence of standardised observable units of analysis in operational audits at the application layer (Mökander et al. 2023b). The AI Epidemiology grammar provides these units of analysis as structured input-output fields, enabling audits to examine where evidential or policy alignment scores degrade, or where override rates differ systematically across models. The grammar therefore complements algorithmic auditing by supplying the per-interaction substrate over which audit standards can be evaluated.

Safety incident reporting systems such as the AI Incident Database (McGregor 2021; Hoffmann and Frase 2025) catalogue salient AI failures retrospectively, drawing on news reports, regulatory filings, and voluntary submissions to build a public record analogous to aviation incident databases. Such records are essential but selective: they record events sufficiently visible to be reported, after the fact, and typically at the level of named incidents rather than routine interactions. The grammar is intended to operate before failure and at the level of routine interactions, enabling it to support triage and intervention before an event becomes a reportable incident. Therefore, while retrospective incident reporting establishes a public catalogue of harms, the grammar establishes a prospective record from which precursors to such harms can be identified at the population level.

## **Differentiating AI Epidemiology from RLHF**

The framework bears a resemblance to reinforcement learning from human feedback (RLHF), insofar as both involve human assessment of AI outputs. However, the two approaches evaluate AI outputs for different purposes at distinct points in the model lifecycle.

The first difference concerns their objectives. RLHF is an alignment technique in which human preference feedback is used to update model parameters so that the model learns to produce outputs humans favour (Christiano et al. 2017; Ouyang et al. 2022). Its goal is to modify model behaviour during training. AI Epidemiology is a governance measurement framework which scores outputs of a deployed model without altering it. RLHF therefore trains a model, whereas AI Epidemiology produces an output-level record of that model's behaviour in deployment.

The second difference concerns signal source and granularity. RLHF relies on preference comparisons or scalar reward signals collected during a training phase, typically from human annotators selecting between paired outputs. AI Epidemiology relies on structured input-output fields (mission, conclusion, justification, risk level) and alignment scores (evidential alignment, policy alignment), and operates across all routine interactions rather than on a curated training sample.

The third difference concerns timing. While RLHF operates before deployment, AI Epidemiology assesses output alignment and risk level during active deployment. This also distinguishes AI Epidemiology from constitutional AI (Bai et al. 2022), direct preference optimisation (Rafailov et al. 2023), and reinforcement learning from AI feedback (RLAIF; Lee et al. 2023), in which AI-generated feedback is used during training rather than as a live governance signal.

## **3 Core components: operationalising AI Epidemiology**

AI Epidemiology requires a structured protocol for systematically categorising expert-AI interactions to enable population-level pattern analysis. We define the grammar as a structured capture protocol comprising six fields: three input-output fields (mission, conclusion, justification), one stratification variable (risk level), and two alignment scores (evidential alignment, policy alignment).

The input-output fields compress each expert-AI interaction into three structured components: what the expert asked (mission), what the AI recommended (conclusion), and why the AI recommended it (justification). All fields are populated automatically from the conversation log without manual data entry from experts to ensure minimum expert burden. Mission, conclusion, and justification are extracted from the conversation log in a first pass; risk level is assigned from the mission type; alignment scores are generated in a

second pass by the LLM judge against applicable reference documents; and override is detected from reprompting within the conversation. Institutions receive both model-agnostic output assessments (e.g. '75% of such outputs were policy-misaligned') and model-specific assessments (e.g. '58% of such GPT-4 outputs were policy-misaligned'). Governance continuity is preserved by making model-agnostic assessments the default until sufficient interaction data accrues for model-specific analysis.

### **Field 1: mission**

Definition: Mission is the task the AI is given, extracted from the expert's input and stripped of conversational scaffolding.

Extraction protocol: The LLM judge extracts mission from the expert's input. A valid extraction specifies: the task type (recommend, assess, classify, draft, or analogous action); the subject (the clinical case, financial instrument, legal matter, or analogous object); and the decision space (the set of actions the institution could take). Verbatim reproduction of conversational text is not a valid extraction. Ambiguous inputs conflating multiple tasks are resolved by capturing multiple missions.

Example (clinical). A free-form prompt: 'I have a 45-year-old patient with a 30 pack-year smoking history presenting with a persistent dry cough for 8 weeks. She has no prior imaging. What should I do?'

Extracted mission (hypothetical): recommend diagnostic workup for persistent dry cough in a 45-year-old patient, with  $\geq 20$  pack-years, symptom duration 8 weeks, prior imaging absent.

### **Field 2: conclusion**

Definition: Conclusion is the AI's recommended action or determination.

Extraction protocol: The LLM judge compresses the AI's output into a structured representation of the primary recommended action, separating any secondary or conditional recommendations. If the AI's output contains no identifiable conclusion, this is itself recorded as a conclusion type.

Example (lending): An AI recommending rejection of a mortgage application.

Extracted conclusion (hypothetical): recommend rejection of mortgage application.

### **Field 3: justification**

Definition: Justification is the evidential basis the AI has invoked for its conclusion, recorded as a category of evidence rather than verbatim text.

Extraction protocol: The LLM judge extracts the categories of evidence the AI has invoked, without assessing whether those claims are correct (as this is the function of the evidential alignment score).

Example (lending): An AI recommending rejection of a mortgage application on the basis that the applicant's credit score of 550 falls below the underwriting threshold.

Extracted justification (hypothetical): appeal to underwriting threshold.

#### **Field 4: risk level**

Definition: Risk level is the stratification variable: it divides the population of expert-AI interactions into high, medium, and low-consequence strata based on the potential severity of harm if the AI's conclusion is incorrect and acted upon.

Scoring rubric: Risk level is assigned from mission type according to pre-specified criteria developed with domain experts before data collection. In clinical settings:

High risk: where an incorrect conclusion could cause serious patient harm, e.g. acute angle closure glaucoma

Medium risk: where harm is recoverable or requires further intervention to detect, e.g. stable ocular hypertension

Low risk: where incorrect conclusions are unlikely to cause direct harm, e.g. dry eye syndrome

#### **Field 5: policy alignment score**

Definition: Policy alignment score measures the degree to which the AI's conclusion is consistent with the policies applicable to the mission type: institutional policy, regulatory policy, or professional guidelines.

Scoring rubric: The LLM judge is provided with the applicable policy document and assesses whether the AI's conclusion is consistent with it.

High: The conclusion is consistent with applicable policy for the mission type. Example (clinical): The AI recommends urgent imaging for a patient with high-risk smoking history and persistent cough, consistent with applicable guidelines for this presentation.

Medium: The conclusion is partially consistent with applicable policy but deviates in a non-material respect, or the policy does not clearly address the mission type. Example (lending): The AI recommends conditional approval where policy endorses approval with compensating factors; the condition imposed is compatible with but more restrictive than policy requires.

Low: The conclusion is inconsistent with or contradicts applicable policy. Example (clinical): The AI recommends watchful waiting for a high-risk patient where applicable guidelines recommend urgent imaging.

### **Field 6: evidential alignment score**

Definition: Evidential alignment score measures the degree to which the factual claims in the AI's justification are supported by the available evidence base for the domain, independently of whether the conclusion conforms to institutional policy. Where policy alignment asks whether the AI's conclusion is institutionally appropriate, evidential alignment asks whether the AI's reasoning is evidentially defensible.

Scoring rubric: The LLM judge checks whether the factual claims in the AI's justification are supported by the available evidence base. Length of justification is not a criterion.

High: The factual claims in the AI's justification are supported by the available evidence base. Example (lending): The AI states that post-bankruptcy recovery patterns are associated with lower default risk; this claim is supported by established credit risk research.

Medium: The factual claims are partially supported but the evidence is incomplete, mixed, or unclear, without definitively contradicting the conclusion. Example (lending): The AI states that self-employment income is a significant default risk factor; the evidence on this relationship is mixed and context-dependent.

Low: The factual claims are unsupported by or inconsistent with the available evidence base. Example (lending): The AI states that a short employment history is a strong independent predictor of mortgage default; the claim is not supported by established credit risk research, which identifies it as a weak predictor when income stability is controlled for.

### **Field 7: override**

Definition: Override is a binary field indicating whether the expert has explicitly contested the AI's original conclusion.

Detection protocol: The LLM judge scores the AI's original output for evidential and policy alignment. Where scores are medium or low, the interface flags the output for expert review. The expert either accepts the output or activates an explicit override mechanism within the conversation interface, ensuring that the override signal reflects deliberate expert contestation rather than conversational continuation or clarification. Override is recorded as yes when the expert activates this mechanism, and no when the expert accepts the output.

Example (lending): The LLM judge flags a mortgage rejection recommendation for low evidential alignment. The expert reviews the output and activates the override mechanism.

Override: yes.

### **Field 8: corrective option**

Definition: Corrective option is the revised conclusion accepted by the expert following an override, captured automatically from the conversation log.

Detection protocol: Where override is yes, the expert may engage in one or more of the following: reprompting the AI until a revised output achieves acceptable alignment; attaching their own sources or evidence to the conversation, which the AI incorporates into a revised output; or providing their own conclusion directly, flagging the AI output as overridden and attaching supporting evidence for compliance purposes. The corrective option is whatever the expert ultimately accepts or provides as the preferred resolution. Where override is no, corrective option is null.

Example (lending): Following the override above, the expert reprompts with "reassess accounting for the applicant's recovery trajectory since the bankruptcy." The AI revises its recommendation to conditional approval. The expert sees this has high evidential alignment and accepts this output.

Corrective option: conditional approval, recovery trajectory acknowledged.

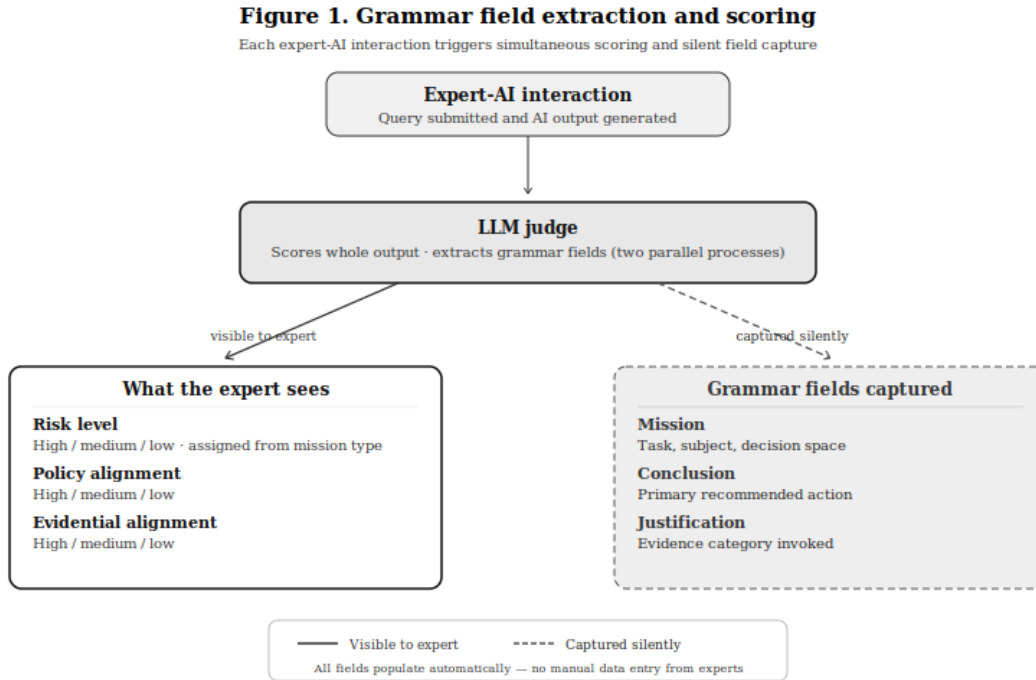
### **The governance rationale for separating the two alignment scores**

The two alignment scores assess the AI's output from complementary perspectives. Policy alignment asks whether the conclusion is institutionally sanctioned; evidential alignment asks whether the factual basis offered for it is defensible. An output can be policy-aligned but evidentially weak, in which case the AI reaches a policy-aligned answer with poor justifications. By contrast, an output can also be evidentially well-grounded but policy-misaligned, in which case the AI may have reasoned competently from sources but reached a recommendation the institution does not endorse. This distinction helps to diagnose output failure more precisely and assess risk more thoroughly.

### **3a. Workflow overview**

The following figures illustrate how the grammar operates in practice.

Figure 1 illustrates how each expert-AI interaction simultaneously produces a scored output visible to the expert and a silent grammar record captured in the background.

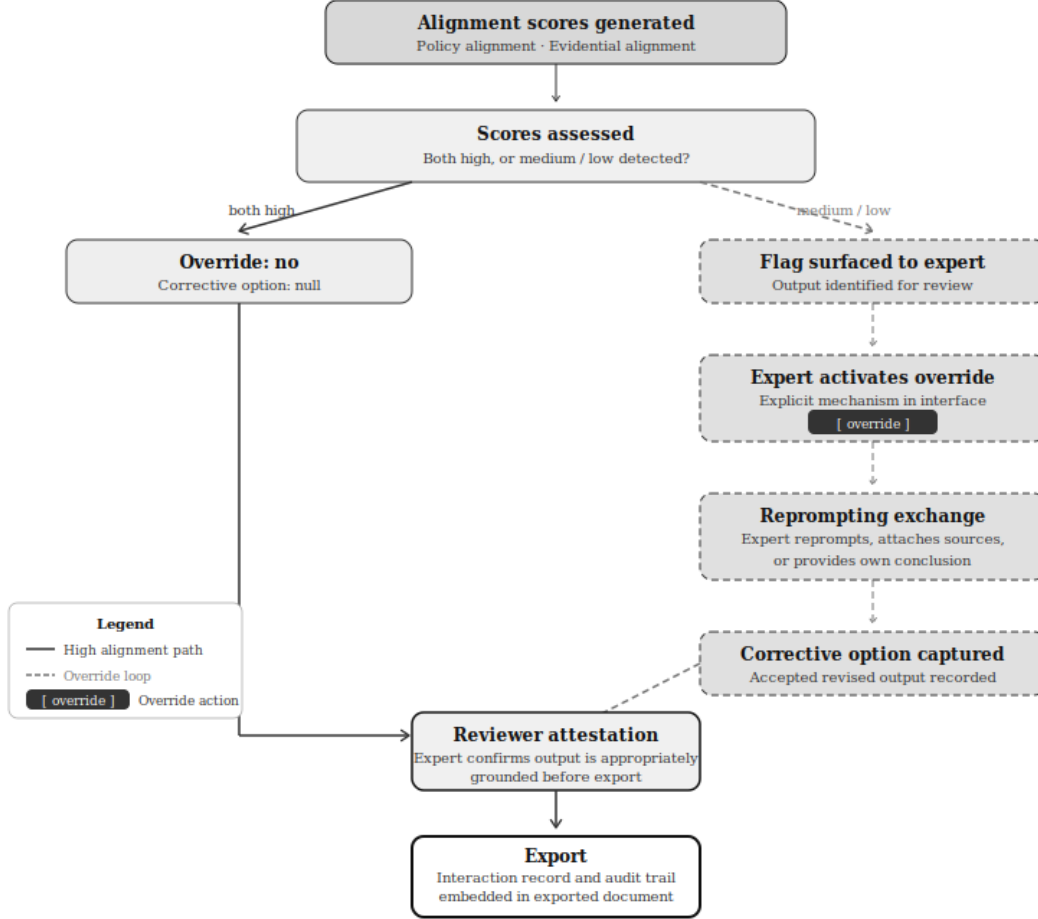


**Figure 1.** Grammar field extraction and scoring. The LLM judge scores the whole AI output against applicable reference documents, producing risk level, policy alignment, and evidential alignment scores visible to the expert. At the same time, it extracts the input-output fields (mission, conclusion, and justification) silently in the background for future population-level analysis. All fields populate automatically without manual data entry from experts beyond their existing interactions.

Figure 2 illustrates the governance decision point that follows scoring. Risk level indicates the consequence severity of the interaction and is used for triage. Where both alignment scores are high, the output proceeds to reviewer attestation and export. Where either score is medium or low, the expert is invited to review, override, and if necessary reprompt until satisfied before exporting.

**Figure 2. Override detection and corrective option capture**

How alignment scores trigger expert review, override is recorded, and the interaction record is completed



**Figure 2.** Override detection and corrective option capture. Both high-scoring outputs and outputs with corrective options require reviewer attestation before export. Medium or low alignment scores trigger expert review via an explicit override mechanism. The corrective option is captured from the reprompting exchange and recorded alongside the original output in the interaction record and audit trail.

#### 4 The LLM-as-judge foundation

Runtime alignment scoring requires automation to achieve scale. Manual expert assessment of every AI output would be impractical and defeat the purpose of a framework designed to minimise expert burden. We therefore propose large language models in the LLM-as-judge role to score policy alignment and evidential alignment automatically. Although there is a structural circularity in employing one black-box model to judge the outputs of another, the LLM-as-judge literature has shown that under bounded conditions, LLM judges achieve reliability comparable to, and in some tasks exceeding, human raters. Three findings bear directly on what bounded conditions achieve.

Firstly, on tasks where criteria are well-specified and inputs are structured, agreement between strong LLM judges and human raters approaches the level typically observed between human raters themselves (Gu et al. 2024). Using MT-Bench and Chatbot Arena, Zheng et al. (2023) report that GPT-4 as a judge achieves over 80% agreement with human preferences on bounded evaluation tasks. Liu et al. (2023) show in G-Eval that structured chain-of-thought scoring with explicit criteria substantially improves alignment with human ratings across summarisation and dialogue quality tasks. Further, Dubois et al. (2024) show that length-controlled win rates improve Spearman correlation with human preferences from 0.94 to 0.98. These findings suggest that bounded conditions consistently improve LLM judge reliability on structured evaluation tasks. The framework introduces bounded conditions through defined rubrics, RAG grounding against applicable reference documents, and low temperature scoring, each designed to satisfy the operating conditions under which reliable LLM judging has been demonstrated.

Secondly, reliability improves when scoring criteria are narrow and score levels are few and explicitly anchored. It degrades when criteria are broad, levels are many and unanchored, and inputs are unstructured conversational text (Li et al. 2024). The grammar's three-level alignment scale and explicit rubric anchors are designed to satisfy these conditions.

Thirdly, however, bounded conditions do not eliminate systematic bias. Three classes are empirically documented and relevant to the framework's scoring design. Sycophancy: LLMs defer to positions asserted in text regardless of correctness (Sharma et al. 2023; Malmqvist 2024; Chen et al. 2025). Self-preference: LLM evaluators recognise and favour outputs from their own model family, a bias that persists under rubric anchoring (Panickssery et al. 2024). Verbosity: judges favour longer responses irrespective of quality (Dubois et al. 2024).

RAG grounding is the primary mitigation for both sycophancy and self-preference: a RAG-anchored judge relies on an external reference corpus rather than its own trained representations, which is the source from which both biases partly originate. Verbosity bias is mitigated by the explicit rubric instruction that length is not a criterion; low temperature reduces random score variation and improves consistency across repeated assessments; and finally, residual bias across all three classes is detected by the reliability verification procedure set out below.

### **The reliability verification procedure**

LLM judge reliability must be empirically verified before scores are used for governance decisions. The procedure comprises three elements.

First, human-judge agreement. Cohen's  $\kappa$  is computed between the LLM judge and domain expert raters on a stratified sample, stratified by domain and risk level. Agreement is

assessed separately by domain, because domain expertise interacts with rubric clarity in ways that may differ across settings. Acceptable thresholds are pre-specified before data collection and calibrated to domain standards: a minimum of  $\kappa \geq 0.60$  consistent with substantial agreement (Landis and Koch 1977), with higher thresholds applied in high-risk strata where the cost of scoring error is greatest.

Second, consistency. The same sample of interactions is scored twice by the same judge configuration and score stability is assessed. A test-retest correlation of 0.70 or above indicates good repeatability (Bland and Altman 1986). Drift of more than one level on the three-point scale between scoring runs is diagnostic of temperature instability or prompt sensitivity.

Third, bias diagnostics. Three targeted tests address the biases identified as relevant to the framework's scoring design. Sycophancy is tested by presenting interactions where the AI output contains strongly assertive language and measuring whether scores drift toward high alignment irrespective of evidential grounding. Empirical benchmarks suggest sycophancy effects of 6-22% are common across LLM evaluation contexts (Panickssery et al. 2024); a drift threshold of 10% is adopted as the pre-specified limit. Verbosity is tested by presenting the same interaction with short and long AI outputs and measuring score drift, assessed by position consistency across two runs (Gu et al. 2024). Self-preference is tested by presenting identical interactions to judges from the same and different model families as the AI being assessed, with score differences between conditions treated as the bias indicator (Wataoka et al. 2024). Acceptable drift thresholds for verbosity and self-preference are pre-specified before data collection.

If pre-specified thresholds are met and bias diagnostics produce no pattern above the pre-specified drift thresholds, the framework proceeds to population-level analysis. If not, the rubric is revised and the procedure repeated until acceptable thresholds are met.

## **5 A specified protocol for empirical evaluation**

We seek to answer whether compressing a raw expert-AI interaction into four structured fields loses too much information to be useful for population-level risk detection. The protocol tests this by comparing how well grammar fields predict alignment scores against how well the full conversational text predicts them. If the structured fields perform nearly as well, the grammar's compression is justified. If not, the grammar needs refinement. Execution of the protocol is left to subsequent empirical work.

### **Design**

The evaluation is specified over a corpus of expert-AI interactions spanning three principal domains: clinical, lending, and legal. The corpus is required to comprise interactions at each of the three policy alignment levels and each of the three evidential alignment levels,

crossed with the three risk levels. This ensures the sample covers the full range of cases the framework would encounter in practice, rather than being skewed toward straightforward interactions. The route to assembling such a corpus from real institutional deployments is set out in the empirical programme of Section 8.

### **Formal statement of the estimand**

The grammar fields for an expert-AI interaction are denoted  $X$ , comprising mission  $M$ , conclusion  $C$ , justification  $J$ , and risk level  $R$ . The policy alignment score is denoted  $Y_p$  and takes values in  $\{0, 1, 2\}$ , and the evidential alignment score  $Y_e$  likewise. We seek to establish whether the four grammar fields hold reliable predictive power over whether the alignment score will be high, medium, or low.

For pre-specified thresholds  $\tau_p$  and  $\tau_e$ , the framework estimates the probability that an interaction's alignment score falls below the threshold given its grammar field values. This is the probability of a misalignment event given the observed interaction characteristics, which is the AI Epidemiology analogue of the conditional risk functions estimated in classical epidemiological surveillance. The estimator is a supervised classifier: a model trained on accumulated grammar field values to learn which combinations of mission, conclusion, justification, and risk level tend to produce low alignment scores. The whole-input comparator estimates the same probability using the full conversational text, providing a performance benchmark. Two interactions are treated as similar to the extent that they share grammar field values and receive similar predicted misalignment probabilities, a definition of similarity that is directly interpretable in governance terms.

### **Uncertainty and sample size**

Furthermore, we need to know how confident we can be in our estimations of how well grammar fields predict alignment scores. We quantify this uncertainty using paired bootstrap inference, a procedure in which the dataset is resampled with replacement 10,000 times and the analysis is run on each resample. The variation in results across these runs produces 95% confidence intervals that characterise the plausible range of the true performance difference between the grammar-field model and the whole-input comparator (Efron and Tibshirani 1993). The method of pairing the bootstrap on interactions ensures that the comparison accounts for the fact that both models are evaluated on the same set of interactions rather than independent samples. A sample size of approximately 500 interactions per domain is sufficient to reliably detect whether the grammar meets the non-inferiority threshold, assuming the models perform at a level typical of clinical risk prediction tasks of comparable complexity. If performance is somewhat lower than expected, approximately 800 interactions per domain would be required.

## **Comparison protocol and non-inferiority margin**

We measure model performance using the area under the receiver operating characteristic curve, commonly called AUC. This statistic captures how well a model separates high-risk from low-risk interactions across all possible thresholds. A perfect model scores 1.0 and a model that performs no better than random chance scores 0.5. We compare the AUC of the grammar-field model against the AUC of the whole-input model using DeLong's test, a statistical test designed for comparing two models evaluated on the same set of interactions (DeLong et al. 1988). Because we run this comparison separately for policy alignment and evidential alignment, we apply the Holm-Bonferroni correction, which adjusts our significance thresholds when multiple tests are run simultaneously to prevent false positives from accumulating (Holm 1979).

Rather than asking whether the grammar-field model is better than the whole-input model, we ask whether it is close enough. We pre-specify a non-inferiority margin  $\delta = 0.05$ , consistent with the standard margin used in diagnostic accuracy studies comparing paired AUC curves.

## **Reviewer effects and temporal drift**

Different judges may score consistently but at slightly different average levels. We account for this by including judge identity as a random effect in the population-level analysis, which separates genuine rubric signal from judge-specific variation (Bates et al. 2015).

Policies and guidelines change over time, and an alignment score calibrated against year-one policy may no longer be accurate in year three. We address this through version-stamping: each scored interaction is linked to the specific version of the policy and evidence corpus applicable at the time of the interaction, ensuring that historical scores are not retroactively altered by subsequent policy changes. We also monitor alignment score distributions over time using rolling-window analysis, which flags structural shifts that may signal model drift or policy change. The reliability verification procedure is re-run periodically on contemporary interaction samples to detect changes in judge behaviour over time.

## **6 Iterative grammar refinement**

The grammar's field definitions are subject to iterative refinement, driven by the feedback signal from the evaluation protocol of Section 5. Where that protocol returns a failing result, field definitions and rubric anchors are revised and the evaluation is repeated. This section specifies how that revision proceeds.

The CDC's Lipids Standardisation Program offers a direct precedent for this refinement procedure. Early cholesterol measurements were inaccurate and inconsistent across

laboratories, containing significant bias that prevented valid population-level comparisons (Hoerger et al. 2011). The CDC's response was iterative refinement of the measurement instruments themselves, including reagents, calibrators, and reference methods, until measurements were sufficiently accurate and comparable for epidemiological use (Cooper et al. 1991). Manufacturers whose reagents performed poorly were required to improve them before certification. Only once that standardisation had been achieved could the Framingham Heart Study establish reliable associations between cholesterol levels and cardiovascular events (Dawber et al. 1951).

The grammar refinement procedure follows the same logic. Its measurement instruments, specifically the field definitions for mission, conclusion, and justification, and the rubric anchors for policy and evidential alignment, are refined until structured-fields prediction meets the non-inferiority criterion. Each refinement cycle is guided by the AUC comparison result from the accumulated corpus of scored interactions, allowing the grammar to be revised in direct response to where structured-fields prediction falls short. Since interactions accumulate automatically as the framework is deployed in institutions, reliable measurement standardisation becomes achievable in a substantially shorter timeframe than traditional epidemiological measurement programmes required.

### **Refinement procedure**

Before refinement begins, the corpus is partitioned into a development split of 80% and a frozen test split of 20%. All refinement decisions are made on the development split only. The frozen test split is reserved for a single evaluation at the end of the refinement procedure, providing an uncontaminated measure of how the final grammar performs on interactions it has never been tuned against.

Refinement proceeds in cycles using the development split. In each cycle, the structured-fields model is estimated and its AUC compared to the whole-input model. Where the non-inferiority criterion is unmet, the grammar is revised. Consider the following example: the structured-fields model consistently underperforms on predicting evidential alignment scores for lending interactions. Inspection reveals that the justification field is capturing 'appeal to credit risk data' as a single category, when in practice some appeals cite specific regulatory guidance and others cite only internal underwriting assumptions. The whole-output model picks up this distinction but the compressed field loses it. The refinement response is to create two separate justification categories: 'appeal to regulatory guidance', and 'appeal to internal underwriting assumption', so the grammar now captures a distinction that was previously invisible to the structured-fields model. In the next cycle, the AUC gap narrows. If it meets the non-inferiority criterion, the refined grammar version proceeds to the frozen test split.

Finally, a parsimony check runs alongside each cycle to guard against adding categories indiscriminately. Where a new field definition does not produce a commensurate improvement in AUC on internal cross-validation, it is rejected. This prevents the grammar from becoming so finely specified that it fits the development data perfectly but fails to generalise. A grammar that performs well on the development split but poorly on the frozen test split is diagnostic of exactly this problem, and the refinement history is re-evaluated to identify where over-specification has occurred.

The refinement procedure therefore ensures the grammar reaches the standard required for prospective risk detection. A grammar that meets the non-inferiority criterion on both the development and frozen test splits produces alignment scores sufficiently reliable for governance use and provides the measurement foundation on which the epidemiological programme set out in Section 8 can build.

## **7 The acknowledged disanalogy and threats to validity**

Although AI Epidemiology is grounded in epidemiological methodology, it differs from traditional epidemiology in ways that create specific challenges for practical deployment. This section acknowledges these differences directly.

The acknowledged disanalogy: constructed outcomes

The most fundamental difference is that our outcome variables are constructed rather than natural. In classical epidemiology, the outcome variable is a directly observable event: a death, a disease diagnosis, a physiological measurement. In AI Epidemiology, the outcome variables are LLM-generated alignment scores. These are not directly observable; they are produced by an automated scoring procedure subject to the biases described in Section 4.

This disanalogy has two implications. First, the validity of the framework depends partly on the validity of the scoring procedure, which must be empirically established through the judge reliability sub-protocol rather than assumed. Second, the framework's prospective risk detection operates at one remove from the harms it ultimately aims to prevent: it detects alignment failures expected to predict expert overrides and downstream adverse outcomes, but the connection to objective downstream harms is not established by the concept paper and must be established by the empirical programme.

### **Threats to validity**

Judge bias. If the judge systematically favours longer justifications, evidential alignment scores will be inflated for verbose outputs and deflated for concise ones, degrading predictive validity. The judge reliability sub-protocol detects this; the rubric anchors mitigate it by specifying that length is not a criterion.

Surveillance bias. In real deployments, not all cases are reviewed equally: high-stakes cases attract disproportionate review. This may produce artificially high apparent misalignment rates in the high-risk stratum, because experts scrutinise those cases more carefully. Stratification by risk level allows analysts to test for this pattern rather than absorbing it into undifferentiated population-level statistics. Operationally, surveillance bias is diagnosed by modelling override probability as a function of grammar fields with review intensity (cases reviewed per stratum per unit time) entered as an interaction term: a significant interaction between stratum and review intensity indicates that apparent misalignment is co-varying with attention rather than with the underlying signal.

Temporal drift. Institutional policies change and guidelines are updated. A grammar calibrated to year-one policy may become systematically misaligned in year three without any change in the AI model. The refinement procedure's cross-validation scheme detects performance degradation over time; sustained degradation triggers rubric re-evaluation.

Confounding by intervention. Once alignment scores are surfaced to experts, review practices may adjust: low-alignment outputs receive more scrutiny, leading to more overrides; high-alignment outputs receive less. The passive monitoring design, in which grammar fields are populated automatically and experts do not see the scores, prevents this confounding during data collection.

Construct validity. Policy alignment score and evidential alignment score, even when reliably scored, may not capture the governance-relevant properties they are intended to measure. An output that scores high on policy alignment may still cause harm if the policy itself is inadequate. Construct validity requires the longitudinal connection to objective downstream outcomes that stage 3 of the empirical programme is designed to establish.

## **8 The three-stage future programme**

AI Epidemiology is operationalised through a three-stage empirical programme forming a complete path from measurement feasibility to prospective risk detection.

The precedent for staged empirical development is the Goldberger pellagra programme. In 1914, Goldberger observed that pellagra, which killed at least 100,000 Americans between 1907 and 1940 (Jarow 2014), was correlated with dietary patterns rather than infectious exposure. Despite lacking mechanistic understanding, his pattern-based interventions were a resounding success: dietary improvements in Mississippi orphanages eliminated pellagra cases, and feeding brewer's yeast to 50,000 Mississippi flood survivors cured thousands and prevented new cases (Jarow 2014; Mooney et al. 2014). The dietary patterns he identified subsequently directed researchers to the missing nutrient: Conrad Elvehjem identified niacin deficiency as the cause in 1937 (Jarow 2014), twenty-three years after Goldberger's initial observations. The three-stage

programme follows the same logic: establish measurement, demonstrate prediction, and ultimately connect predicted risks to observed outcomes.

### **Stage 1: privacy-preserving risk triage (claim a)**

Stage 1 deploys the grammar across institutional partner sites with privacy-preserving infrastructure. The primary objective is to establish that measurement standardisation is achievable at scale in real institutional environments: that independent scorers, or suitably configured LLM judges following the rubrics, produce consistent policy alignment and evidential alignment scores across cases, domains, and institutions. The target is an intraclass correlation coefficient (ICC) of at least 0.70 for each outcome variable, representing good reliability suitable for epidemiological use (Koo and Li 2016).

Privacy-preserving infrastructure ensures that raw conversational data remains solely with the deploying institution. The grammar's semantic compression addresses this requirement: structured mission, conclusion, justification, risk level, and outcome score values can be shared across institutions for cross-institutional analysis without sharing the underlying conversational text. This is both a privacy protection and a demonstration that the grammar's abstraction is sufficient for population-level analysis.

Commercially, the grammar delivers a comprehensive audit trail from deployment by documenting every AI-institution interaction to satisfy regulatory compliance requirements. This ensures that institutions gain governance value before population-level reliability patterns emerge, addressing the bootstrap challenge that foundational epidemiological studies operated under academic timelines while AI Epidemiology must present value to institutions within months.

Stage 1 corresponds to claim (a): it tests whether the measurement instrument produces reliable output under real deployment conditions and establishes that the circularity is bounded. Its successful completion establishes that the judge applies the rubric consistently against the reference documents, and that different judges applying the same rubric to the same interaction reach consistent conclusions.

### **Stage 2: prediction of expert overrides (claim b)**

Stage 2 tests whether grammar fields predict expert override behaviour at the individual interaction level. The primary analysis estimates the association between grammar fields and expert override using logistic regression with Holm-Bonferroni correction. DeLong's test on paired AUCs compares structured-fields prediction to whole-input prediction, with the pre-specified non-inferiority margin  $\delta = 0.03$ .

Stage 2 also tests whether exposure-outcome associations are consistent across risk level strata or vary by stratum. Strong associations in the high-stakes stratum and weak

associations in the low-stakes stratum would suggest that alignment scoring is most valuable where the consequences of error are greatest, which is the expected pattern and which can be tested by stratified Holm-Bonferroni analyses.

Stage 2 corresponds to claim (b): it tests whether alignment scores predict the external validator (expert override). A positive result establishes that alignment scores track something human experts recognise as misalignment. A negative result calls for rubric revision rather than abandonment of the framework: it indicates the measurement instrument is not yet calibrated to what experts recognise as misalignment.

### **Stage 3: true epidemiology (claim c)**

Stage 3 connects the framework to objective downstream outcomes in industries where such outcomes are observable: loan default in lending, case outcome or regulatory sanction in legal practice, patient adverse event or mortality in medicine. This is the AI Epidemiology analogue of outcome tracking in clinical epidemiology.

Stage 3 is the critical test of construct validity. If alignment scores predict expert overrides (stage 2) and expert overrides predict downstream adverse outcomes (a condition stage 3 tests directly), the framework has demonstrated prospective risk detection from observable grammar fields to objective harms. The methodological challenge is confounding by mediation: adverse outcomes have many causes, and isolating the contribution of the AI interaction's grammar fields requires a design that can distinguish these contributions, such as an instrumental variable design where variation in AI output quality is exogenous to expert behaviour.

Stage 3 also provides the data necessary to direct mechanistic interpretability research in the way that Goldberger's dietary pattern observations guided the search for niacin: by identifying specific failure patterns documented to cause harm, it directs researchers in mechanistic interpretability to examine the corresponding circuits; in SHAP and LIME to analyse feature importance in documented failures; and in chain-of-thought research to investigate why models cannot articulate the relevant distinctions. The governance goal and the scientific goal are therefore served simultaneously by the same structured data.

## **9 Domain of applicability**

The framework's domain of applicability should be stated explicitly rather than left to be inferred from the examples. This section provides that statement.

### **In scope**

The framework is designed for high-stakes institutional decision-making settings in which AI outputs inform decisions made by accountable human reviewers against an applicable policy and evidence base. Three conditions jointly define in-scope settings. First, there must be an applicable policy corpus: a body of institutional policy, regulatory guidance, or professional guidelines that specifies what conclusions are appropriate for the mission type. Second, there must be an applicable evidence base: a domain with established expert practice supported by an identifiable corpus of authoritative sources. Third, there must be accountable human reviewers who can override AI outputs and whose override behaviour is observable.

Settings meeting these conditions include clinical practice (medicine, nursing, pharmacy), regulated financial services (lending, insurance underwriting, investment advice), legal practice in regulated jurisdictions, and regulated public-sector decision-making (welfare eligibility, planning decisions, regulatory enforcement). These settings share the structural property that AI outputs are not the final decision: a human with professional accountability reviews the AI's recommendation and takes or overrides it.

### **Out of scope**

Four categories of setting fall outside the framework's scope, and we state this directly.

Open-ended creative tasks (content generation, creative writing, coding assistance without a specified correctness criterion) have no applicable policy corpus and no defined evidence base. Policy alignment and evidential alignment scores are not meaningful when there is no policy to align with and no evidence to be grounded in.

Low-stakes consumer applications, where the potential harm from an incorrect AI output is below the threshold that justifies governance overhead, do not benefit from the framework's measurement infrastructure. The framework's deployment cost (judge infrastructure, taxonomy development, expert reviewer time) is justified only where the risk of misalignment is consequential.

Applications where there is no reviewable expert practice against which alignment can be defined cannot be in scope. The framework's validation pathway depends on the existence of human experts whose override behaviour provides an external referent for the alignment scores; where no such expert practice exists, the validation chain is broken.

Real-time interactive AI without expert review at the cadence of interaction (high-frequency trading, real-time recommender systems, interactive chatbots without human-in-the-loop review) falls outside the framework's governance model as specified. The measurement layer can be applied asynchronously to such systems, but the override-validation step in claim (b) requires reviewable expert practice and has no foothold where review is structurally absent.

The framework provides a measurement instrument for those settings in which population-level governance is required because the stakes of misalignment are consequential, human reviewers are present and accountable, and a policy and evidence base against which alignment can be assessed exists, rather than a universal claim that interpretability or governance is required for all AI applications. These conditions are jointly necessary; where any one is absent, the framework falls outside its domain of applicability.

The in-scope conditions carry institutional politics. Authorship of the applicable policy corpus is itself an institutional question: the corpus reflects the priorities of whichever body issues guidelines, and where multiple bodies issue conflicting guidance the framework's three-source hierarchy resolves the conflict procedurally while leaving the underlying contest over authority to be settled elsewhere. Jurisdictional variation in policy is an empirical pattern that mission-stratified analysis would surface, rather than a complication for the framework: the same mission type, scored against different jurisdictional policy corpora, will produce different alignment distributions, and that variation is itself a governance-relevant signal. The framework therefore presupposes the institutional conditions of high-stakes professional practice and inherits the contested politics of those conditions.

## **10 Conclusion**

AI Epidemiology represents a fundamental shift in how institutions govern and explain AI systems at scale. Instead of attempting to understand model internal computations, AI Epidemiology takes the structured interaction as its unit of analysis and seeks to stratify and predict the risk of each output through population-level patterns.

The framework advances three claims in sequence. Claim (a) is a measurement instrument claim: LLMs, under bounded conditions, can produce reliable assessments of policy alignment and evidential alignment. The grammar's field definitions, extraction protocols, and scoring rubrics are designed to satisfy the operating conditions under which the LLM-as-judge literature reports reliable automated scoring. Claim (b) is a validation-against-human-oversight claim: LLM-judged alignment scores can be validated against expert override patterns. Claim (c) is a validation-against-outcomes claim: alignment scores can be mapped against downstream objective outcomes, constituting epidemiology in the Bradford Hill sense.

The structural circularity in using a black-box judge to evaluate black-box outputs is acknowledged and not minimised. Bounded conditions reduce the circularity by constraining the judge to a rubric and grounding it in reference documents; they do not break the circularity, because systematic biases persist. The three-stage empirical

programme progressively reduces the circularity by adding external referents at each stage.

The framework's contribution is risk detection at the population level: structured, human-comprehensible patterns of judgment about AI interactions that institutions can act on without access to model internals. This paper is towards AI Epidemiology because the full Bradford Hill standard requires claim (c), which the empirical programme has not yet established. Each stage adds an external referent and reduces the residual circularity. We are not yet at full epidemiology; we are moving toward it.

In the tradition of Bradford Hill and Goldberger, AI Epidemiology pursues prospective intervention on the basis of observable associations, with the expectation that mechanistic understanding will follow as researchers use the framework's documented failure patterns to guide targeted interpretability research. Both the governance goal and the scientific goal are served: the institution gains prospective risk detection now, and the research community gains structured failure data to drive mechanistic discovery over the longer term.

## 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](https://scholarship.law.uwyo.edu/cgi/viewcontent.cgi?article=1144&context=faculty_articles)

[i/viewcontent.cgi?article=1144&context=faculty\\_articles](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>

Bai Y, Jones A, Ndousse K, Askell A, Chen A, DasSarma N, Drain D, Fort S, Ganguli D, Henighan T, Johnston N, Joseph N, Kadavath S, Kernion J, Conerly T, El Showk S, Elhage N, Hatfield-Dodds Z, Hernandez D, Hume T, Johnston S, Kravec S, Lovitt L, Nanda N, Olsson C, Amodei D, Brown T, Clark J, McCandlish S, Olah C, Mann B, Kaplan J (2022) Constitutional AI:

harmlessness from AI feedback. arXiv preprint arXiv:2212.08073  
[https://doi.org/10.48550/arXiv.2](https://doi.org/10.48550/arXiv.2212.08073)

212.08073

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>

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>

Christiano P, Leike J, Brown T, Martic M, Legg S, Amodei D (2017) Deep reinforcement learning

from human preferences. In: Proc Adv Neural Inf Process Syst 30. Curran Associates, Inc. <http>

[s://doi.org/10.48550/arXiv.1706.03741](https://doi.org/10.48550/arXiv.1706.03741)

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/>

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>

DeLong ER, DeLong DM, Clarke-Pearson DL (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics

44(3):837-845 <https://doi.org/10.2307/2531595>

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.38142.5544>

79.AE

Dubois Y, Galambosi B, Liang P, Hashimoto TB (2024) Length-controlled AlpacaEval: a simple

way to debias automatic evaluators. arXiv preprint arXiv:2404.04475  
[https://doi.org/10.48550/ar](https://doi.org/10.48550/arXiv.2404.04475)

Xiv.2404.04475

Fletcher GS, Fletcher SW, Fletcher RH (2020) Clinical epidemiology: the essentials. 6th edn. Philadelphia PA: Wolters Kluwer/Lippincott Williams and Wilkins. ISBN 9781975109554

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

Gu J, Jiang A, Zhu X, Dong J, Zeng Z, Jia L, He X, Ding Y, Wu H (2024) A survey on LLM-as-a-

judge. arXiv preprint arXiv:2411.15594 <https://doi.org/10.48550/arXiv.2411.15594>

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>

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.](https://www.cdc.gov/pcd/issues/2011/nov/10_0253.htm)

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>

Horovitz 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/](https://doi.org/10.48550/arXiv.2407.10114)

[g/10.48550/arXiv.2407.10114](https://doi.org/10.48550/arXiv.2407.10114)

Jarrow G (2014) Red madness: how a medical mystery changed what we eat. Honesdale PA:

Calkins Creek <https://astrapublishinghouse.com/product/red-madness-9781590787328/>

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  
<https://doi.org/10.48550/arXiv.2205.15419>

[0/arXiv.2205.15419](https://doi.org/10.48550/arXiv.2205.15419)

Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data.

Biometrics 33(1):159-174 <https://doi.org/10.2307/2529310>

Lee H, Phatale S, Mansoor H, Lu K, Mesnard T, Bishop C, Carbune V, Rastogi A (2023) RLAIIF: scaling reinforcement learning from human feedback with AI feedback. arXiv preprint

[arXiv:2309.00267](https://arxiv.org/abs/2309.00267) <https://doi.org/10.48550/arXiv.2309.00267>

Lee TA, Pickard AS (2013) Exposure definition and measurement. In: Velentgas P et al. (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/NBK1261>

91/

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>

0.48550/arXiv.2005.11401

Li J, Sun M, Han S, Shi C, Luo J (2024) Leveraging large language models for NLG evaluation: a

survey. arXiv preprint arXiv:2401.07103 <https://doi.org/10.48550/arXiv.2401.07103>

Liu Y, Iter D, Xu Y, Wang S, Xu R, Zhu C (2023) G-eval: NLG evaluation using GPT-4 with better

human alignment. arXiv preprint arXiv:2303.16634  
<https://doi.org/10.48550/arXiv.2303.16634>

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 [http](http://doi.org/10.1016/j.jclinepi.2013.08.015)

[s://doi.org/10.1016/j.jclinepi.2013.08.015](http://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>

2307.15771

McGregor S (2021) Preventing repeated real world AI failures by cataloging incidents: the AI

incident database. In: Proc 35th AAAI Conf Artif Intell, 35(17):15458-15463  
<https://doi.org/10.1609/aaai.v35i17.17817>

9/aaai.v35i17.17817

Mökander J, Morley J, Taddeo M, Floridi L (2023) Auditing AI systems: a practical guidance.

Minds and Machines 33:1-17 <https://doi.org/10.1007/s11023-023-09644-0>

Mooney SJ, Knox J, Morabia A (2014) The Thompson-McFadden Commission and Joseph Goldberger: contrasting two historical investigations of pellagra. *Am J Epidemiol* 180(3):235-

244 <https://doi.org/10.1093/aje/kwu134>

Mosqueira-Rey E, Hernandez-Pereira E, Alonso-Rios D, Bobes-Bascarán J, Fernandez-Leal A (2023) Human-in-the-loop machine learning: a state of the art. *Artif Intell Rev* 56(4):3005-3054

<https://doi.org/10.1007/s10462-022-10246-0>

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>

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>

Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, Zhang C, Agarwal S, Slama K, Ray A, Schulman J (2022) Training language models to follow instructions with human

feedback. In: Proc Adv Neural Inf Process Syst 35:27730-27744  
<https://doi.org/10.48550/arXiv.22>

03.02155

Panickssery A, Bowman SR, Feng S (2024) LLM evaluators recognize and favor their own

generations. arXiv preprint arXiv:2404.13076 <https://doi.org/10.48550/arXiv.2404.13076>

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>

Raji ID, Smart A, White RN, Mitchell M, Gebru T, Hutchinson B, Smith-Loud J, Theron D, Barnes P (2020) Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In: Proc 2020 Conf Fairness Accountability Transparency, 33-44

<https://doi.org/10.1145/3351095.3372873>

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>

<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-97801](https://global.oup.com/academic/product/epidemiology-9780197751541)

[97751541](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>

Sharma M, Tong M, Korbak T, Duvenaud D, Askill A, Bowman SR, Cheng N, Durmus E, Hatfield-Dodds Z, Irving G, Kundu S (2023) Towards understanding sycophancy in language

models. arXiv preprint arXiv:2310.13548 <https://doi.org/10.48550/arXiv.2310.13548>

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.2501.1>

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.](https://www.unav.edu/documents/16089811/16155256/Smoking+and+Health+the+Surgeon+General+Report+1964.pdf)

[edu/documents/16089811/16155256/Smoking+and+Health+the+Surgeon+General+Report+1964.](https://www.unav.edu/documents/16089811/16155256/Smoking+and+Health+the+Surgeon+General+Report+1964.pdf)

pdf

Wang P, Li L, Chen J, Zhu D, Lin Z, Cao Y, Liu T, Sui Z, Shi S (2023) Large language models are

not robust multiple choice selectors. arXiv preprint arXiv:2309.03882  
[https://doi.org/10.48550/a](https://doi.org/10.48550/arXiv.2309.03882)

[rXiv.2309.03882](https://doi.org/10.48550/arXiv.2309.03882)

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>

Wu T, Terry M, Cai CJ (2022) AI chains: transparent and controllable human-AI interaction by chaining large language model prompts. In: Proc 2022 CHI Conf Human Factors Comput Syst,

Article 385 <https://doi.org/10.1145/3491102.3517582>

Zheng L, Chiang WL, Sheng Y, Zhuang S, Wu Z, Zhuang Y, Lin Z, Li Z, Li D, Xing E, Zhang H, Gonzalez JE, Stoica I (2023) Judging LLM-as-a-judge with MT-bench and chatbot arena. In:

Proc Adv Neural Inf Process Syst 36 <https://doi.org/10.48550/arXiv.2306.05685>

Anderljung M, Barnhart J, Korinek A, Leung J, O’Keefe C, Whittlestone J, et al. (2023) Frontier AI regulation: managing emerging risks to public safety. arXiv preprint arXiv:2307.03718 <https://doi.org/10.48550/arXiv.2307.03718>

Bommasani R, Hudson DA, Adeli E, et al. (2021) On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 <https://doi.org/10.48550/arXiv.2108.07258>

Burrell J (2016) How the machine ‘thinks’: understanding opacity in machine learning algorithms. Big Data & Society 3(1):2053951715620969 <https://doi.org/10.1177/2053951715620969>

Casper S, Davies X, Shi C, et al. (2023) Open problems and fundamental limitations of reinforcement learning from human feedback. Transactions on Machine Learning Research arXiv:2307.15217

Chen X, Ali SR, Allard CB, et al. (2025) When helpfulness backfires: LLMs and the risk of false medical information due to sycophantic behavior. npj Digital Medicine 8:1147 <https://doi.org/10.1038/s41746-025-02008-z>

Doshi-Velez F, Kim B (2017) Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 <https://doi.org/10.48550/arXiv.1702.08608>

Gao Y, Xiong Y, Gao X, Jia K, Pan J, Bi Y, Dai Y, Sun J, Wang M, Wang H (2024) Retrieval-augmented generation for large language models: a survey. arXiv preprint arXiv:2312.10997 <https://doi.org/10.48550/arXiv.2312.10997>

Gu J, Jiang A, Zhu X, Dong J, Zeng Z, Jia L, He X, Ding Y, Wu H (2024) A survey on LLM-as-a-judge. arXiv preprint arXiv:2411.15594 <https://doi.org/10.48550/arXiv.2411.15594>

Hoffmann M, Frase H (2025) AI incidents: key components for a mandatory reporting regime. Center for Security and Emerging Technology (CSET) Issue Brief <https://doi.org/10.51593/20240023>

Malmqvist L (2024) Sycophancy in large language models: causes and mitigations. arXiv preprint arXiv:2411.15287 <https://doi.org/10.48550/arXiv.2411.15287>

Mökander J, Schuett J, Kirk HR, Floridi L (2023b) Auditing large language models: a three-layered approach. *AI and Ethics* <https://doi.org/10.1007/s43681-023-00289-2>

Mueller A, Jenner J, Bhatt U, et al. (2025) MIB: a mechanistic interpretability benchmark. In: *Proceedings of the 42nd International Conference on Machine Learning*, PMLR 267:45069–45108

Rafailov R, Sharma A, Mitchell E, Ermon S, Manning CD, Finn C (2023) Direct preference optimization: your language model is secretly a reward model. In: *Proc Adv Neural Inf Process Syst 36* arXiv:2305.18290 <https://doi.org/10.48550/arXiv.2305.18290>

Shi L, Ma C, Liang W, Diao X, Ma W, Vosoughi S (2024) Judging the judges: a systematic study of position bias in LLM-as-a-judge. *arXiv preprint* arXiv:2406.07791 <https://doi.org/10.48550/arXiv.2406.07791>

United Nations High-Level Advisory Body on AI (2024) *Governing AI for humanity: final report*. New York: United Nations [https://www.un.org/sites/un2.un.org/files/governing\\_ai\\_for\\_humanity\\_final\\_report\\_en.pdf](https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_en.pdf)