Research Profile Summary - Aman Chandra H

My research spans two interconnected themes: making clinical AI safer and more trustworthy before it reaches patients, and broadening AI's utility in domains where reliable, grounded, and consistent behaviour is essential. The core of my work, five papers, forms a coherent programme in clinical AI evaluation and deployment. Two additional papers extend this into legal AI and multilingual model behaviour.

The clinical AI work begins with a foundational observation: the standard metric used to evaluate ICU mortality prediction models, the F1 score, treats all errors as equally costly, which is clinically indefensible. A missed death is not equivalent to a false alarm. My first paper introduces the Clinical Risk-Weighted Score (CRWS), a new evaluation metric that penalises missed deaths more heavily, and demonstrates on the MIMIC-IV database (65,366 patients) that model rankings shift meaningfully under this clinically honest lens, including exposing an accuracy-safety paradox where the most accurate model misses the most deaths.

The second paper extends this to patient severity. The Risk-Weighted Evaluation Metric (RWEM) incorporates SOFA-based clinical severity directly into the error scoring, so that missing a critically ill patient costs more than missing a lower-acuity patient. Validated across three independent ICU datasets (MIMIC-IV, MIMIC-III, eICU), RWEM reveals consistent model ranking inversions invisible to standard metrics, and introduces a severity-aware threshold strategy that reduces missed deaths by 26.2% without any model retraining.

The third paper turns to a complementary failure mode, probability calibration, asking not just whether models classify correctly, but whether the predicted probability numbers themselves are reliable for high-severity patients. The Severity-Weighted Calibration Error (SWCE) is introduced, and a calibration paradox is demonstrated: the model with the best aggregate calibration score simultaneously exhibits the worst severe-band underestimation, predicting 46% mortality when the true rate is 63.8%, and missing 43 of 69 severely ill patients at a standard triage threshold. This paper also proves mathematically that no aggregate metric can surface subgroup failures when the critically ill constitute only 0.5% of patients; band-stratified analysis is not merely preferable but mathematically necessary.

The fourth paper shifts to deployment, examining what happens when a model trained at one hospital is used at another. Training on MIMIC-IV and validating on eICU (200,234 patients, 208 hospitals), the paper documents that discrimination (AUROC) degrades modestly while calibration collapses approximately nine-fold, and that lightweight post-hoc recalibration fully restores probability reliability without any model retraining, while SHAP analysis identifies which specific features signal distributional drift.

The fifth paper identifies the primary mechanistic driver of that calibration collapse: hospital-level outcome prevalence mismatch. Across 185 real hospitals, calibration error correlates strongly with prevalence gap ($r = 0.52$ to 0.75) while AUROC does not, and a closed-form Bayesian correction requiring only the local mortality rate, with no labeled data and no retraining, reduces calibration error by 55 to 65%, matching the performance of supervised recalibration methods. Together, these five papers constitute a systematic programme: identifying the right metrics, exposing severity-specific failures, and providing deployment-ready corrections for clinical AI systems.

Two parallel papers demonstrate breadth beyond clinical AI. NOMOS is a full-stack Retrieval-Augmented Generation system for Indian legal question answering, grounding a locally deployed Llama 3.2 model in a curated corpus of ten major Indian statutes, including the Bharatiya Nyaya Sanhita 2023, absent from all general-purpose LLM training data, and delivering source-attributed, privacy-preserving answers that substantially outperform general LLMs on citation accuracy, jurisdictional specificity, and hallucination resistance.

The multilingual LLM paper isolates language as an experimental variable across 360 responses from two models, three languages, and three prompt categories, introducing the Behavioral Divergence Score and demonstrating that cross-language divergence consistently exceeds within-language noise. Normative prompts exhibit up to 30% refusal class mismatch across languages, and divergence is model-specific rather than prompt-universal, with direct implications for how multilingual AI safety should be evaluated and audited.

My earliest published work, a Springer-proceedings paper on an IoT-based intravenous fluid monitoring prototype, established the patient safety motivation that runs through all subsequent research: the recognition that gaps in clinical monitoring, whether at the hardware sensor layer or the software evaluation layer, carry direct consequences for patients, and that principled engineering can close them.

For any paper preprint or published version please mail-to : *amanchandrah@gmail.com*

Regards Aman Chandra H

Intravenous Bag Monitoring and Alert System - A Prototype

Published in: Springer CCIS 2635, ICBEST 2025, pp. 560-571

BACKGROUND AND MOTIVATION

Intravenous fluid administration is one of the most routine and critical procedures in healthcare. Nurses and doctors must monitor IV bag levels continuously to ensure patients receive the correct fluid volume and to prevent the line from running dry, an event that can introduce air into the bloodstream, cause electrolyte imbalance, or in severe cases trigger cardiac events. The global error rate associated with IV medication administration is estimated between 9.4% and 97.7%, and approximately 15 to 20% of patients receiving IV treatment in hospitals experience complications linked to inadequate monitoring. In busy clinical settings, continuous manual observation of IV bags is neither practical nor reliable.

PROPOSED SYSTEM

This paper presents a prototype IV Bag Monitoring and Alert System that automates fluid level detection and delivers real-time alerts to healthcare professionals without requiring any contact with the bag or its contents.

The core sensing mechanism is a non-contact capacitive fluid level sensor mounted externally on the IV bag via a Velcro band. Capacitive sensors detect changes in the dielectric properties of the surrounding medium, in this case the presence or absence of fluid behind the bag wall, allowing accurate level measurement without puncturing, touching, or contaminating the bag in any way. This non-contact design makes the device compatible with any standard IV container geometry and eliminates the risk of sensor-introduced contamination.

The sensor is connected to an ESP32 microcontroller, which continuously reads the fluid level, processes the signal, and triggers alerts when the level drops below a predefined threshold. The alert mechanism operates on two channels: a local audible beep signal at the bedside, and a push notification delivered to healthcare staff through the Blynk platform, a cloud-connected IoT application accessible via mobile or web interface. A centralised web-based dashboard allows nursing staff to monitor multiple patients' IV bags simultaneously from a single station, enabling remote surveillance without physical presence at each bedside.

RESULTS AND SIGNIFICANCE

The prototype successfully detects low fluid levels in real time and delivers alerts through both local and remote channels. The system is compact, low-cost, and non-invasive, requiring no modification to the IV bag or administration line. Its compatibility with various container types and its wireless connectivity make it suitable for deployment across diverse clinical environments including general wards, ICUs, and outpatient infusion centres.

The primary contribution is the demonstration that a reliable, low-cost, non-contact IoT-based solution can replace continuous manual monitoring for IV fluid management. By automating the detection and alert process, the system reduces the burden on nursing staff, minimises the risk of human error, and improves patient safety outcomes. The integration with a centralised platform also enables scalable multi-patient monitoring, which is particularly valuable in high-volume or understaffed clinical settings.

SIGNIFICANCE IN THE BROADER RESEARCH PORTFOLIO

This work represents the earliest published contribution in this research programme and establishes the foundational motivation that runs through all subsequent work: the application of technology to address gaps in clinical monitoring and patient safety. While the later papers address software-level failures in clinical AI, how models are evaluated, calibrated, and deployed, this paper addresses the hardware layer, demonstrating that patient safety improvements can be achieved through thoughtful sensor design and real-time alerting, even with lightweight and inexpensive components.

Clinical Risk-Weighted Evaluation of ICU Mortality Prediction Models on MIMIC-IV

BACKGROUND AND MOTIVATION

Machine learning models deployed in ICU settings for mortality prediction are conventionally evaluated using the F1 score, which treats false negatives and false positives as equally costly errors. In clinical practice, this assumption is fundamentally flawed. A false negative in ICU mortality prediction represents a patient predicted to survive who subsequently dies, a missed opportunity for life-saving intervention. A false positive, by contrast, triggers an unnecessary clinical review, which is resource-intensive but rarely fatal. Despite this well-established asymmetry in error costs, virtually no published study operationalises it at the evaluation metric level, and standard metrics continue to obscure clinically dangerous model behaviour, particularly under class imbalance.

PROPOSED METRIC

This paper introduces the Clinical Risk-Weighted Score (CRWS), a confusion-matrix-level evaluation metric that directly encodes the asymmetric cost of false negative and false positive errors through tunable penalty weights: wFN for missed deaths and wFP for false alarms. The metric is grounded in the F-beta family and reduces exactly to the standard F1 score when both weights are set to one, establishing it as a strict generalisation rather than a replacement. For ICU mortality prediction, the authors adopt wFN = 3 and wFP = 1, reflecting a clinical judgment that a missed death warrants three times the penalty of a false alarm. The weights are interpretable to clinical stakeholders in a way that abstract statistical parameters are not, and can be adapted to other tasks such as sepsis detection or readmission prediction by adjusting the cost ratio accordingly.

EXPERIMENTAL SETUP

The study is conducted on MIMIC-IV v3.1, a publicly available critical care database spanning 65,366 unique adult ICU patients with a mortality rate of 10.84%. The prediction target is in-hospital mortality. Four classifiers are evaluated: Logistic Regression, Random Forest, XGBoost, and MLP Neural Network. Ten static tabular features are used, including ICU length of stay, age, admission type, and care unit. Class imbalance is addressed using SMOTE applied exclusively to the training partition. The evaluation protocol includes accuracy, precision, recall, F1, F2, CRWS, MCC, balanced accuracy, AUPRC, and full confusion matrix counts.

RESULTS

The central finding is a pronounced divergence between model rankings under F1 and under CRWS. Random Forest achieves the highest accuracy among all models at 87.23%, yet produces 1,220 missed deaths out of 1,417 actual deaths in the test set, an 86.1% miss rate. Under CRWS, it ranks last with a score of 0.147. XGBoost, with the lowest accuracy at 59.32%, misses only 289 deaths and achieves the highest recall at 0.796, ranking first under CRWS with a score of 0.597.

Model rankings under F1: XGBoost > MLP > LR > Random Forest
Model rankings under CRWS: XGBoost > LR > MLP > Random Forest

The shift in ranking between LR and MLP reflects LR's superior sensitivity under risk-weighted evaluation, a distinction that F1 fails to surface. Bootstrap confidence intervals (95%, n = 500) confirm that CRWS rankings are statistically distinguishable across all model pairs, whereas F1 intervals for LR and MLP partially overlap, rendering them indistinguishable under standard evaluation.

Threshold optimisation guided by CRWS rather than F1 yields substantial reductions in missed deaths across all models. Random Forest reduces its false negative count from 1,220 to 188, an 84.4% reduction, by lowering the decision boundary to favour sensitivity. Sensitivity analysis across wFN values of 1, 2, 3, and 5 confirms that CRWS rankings remain stable beyond the F1-equivalent baseline. An ablation study comparing SMOTE against class-weighting demonstrates that SMOTE consistently produces higher CRWS scores and fewer missed deaths, even in cases where class-weighting yields a marginally higher F1.

CONTRIBUTION AND SIGNIFICANCE

The paper makes a methodological contribution to clinical AI evaluation. It demonstrates that a model can simultaneously achieve the highest accuracy and the worst clinical safety profile, a phenomenon the authors term the accuracy-safety paradox. CRWS provides a principled, mathematically grounded mechanism to detect and penalise such behaviour. The metric is positioned not as a universal replacement for F1 but as a necessary complement in any evaluation pipeline where minimising high-consequence errors is the primary operational objective. Recommended parameter settings are provided for common clinical prediction tasks, derived from the clinical cost-ratio literature.

LIMITATIONS

The study is restricted to a single dataset, a single prediction task, ten static tabular features, and one primary cost assumption. Future work will extend CRWS to multi-class settings, incorporate clinical severity scores such as SOFA and APACHE as adaptive penalty weights, and validate the metric across additional tasks including sepsis onset detection and unplanned hospital readmission.

Severity-aware Evaluation of ICU Mortality Prediction Models using Risk-Weighted Error Metrics

NOTE

This paper is a direct and substantially extended continuation of the preceding work on the Clinical Risk-Weighted Score (CRWS). While the earlier paper established the foundational argument that standard metrics fail to account for asymmetric error costs in ICU mortality prediction, this work advances that framework by incorporating patient-level clinical severity into the evaluation metric itself, validating across three independent datasets and five model classes, and demonstrating a practical deployment strategy that improves clinical safety without any model retraining.

BACKGROUND AND MOTIVATION

Standard evaluation metrics for ICU mortality prediction, including F1-score and AUC-ROC, assess model performance under the implicit assumption that all prediction errors carry equal clinical consequence. This assumption is flawed in two respects. First, a missed death is far more consequential than a false alarm. Second, and less commonly recognised, not all missed deaths are equally consequential: failing to identify a patient with a SOFA score of 19 (mortality risk exceeding 70%) represents a substantially greater clinical failure than missing a patient with a SOFA score of 2 (mortality risk below 10%). No widely adopted evaluation framework accounts for this patient-level severity dimension, and none can be applied retrospectively to an already-trained model without modifying its training pipeline.

PROPOSED FRAMEWORK

This paper introduces the Risk-Weighted Evaluation Metric (RWEM), a severity-aware evaluation framework that incorporates patient acuity, measured via the Sequential Organ Failure Assessment (SOFA) score, directly into the error scoring at evaluation time. RWEM assigns a higher penalty to missed deaths among critically ill patients and a lower penalty to missed deaths among lower-severity patients, scaled continuously by the patient's normalised SOFA score.

A tunable parameter alpha controls the degree of severity sensitivity. When alpha is set to zero, RWEM reduces to a standard asymmetric cost-sensitive misclassification rate. When alpha is set to one, the primary experimental setting, severity weighting is fully active. The cost ratio between false negatives and false positives is set at 10:1, grounded in published economic evidence on the differential costs of missed high-severity ICU cases. Critically, RWEM requires no modification to any trained model; it is applied exclusively at evaluation time, making it immediately applicable to any existing clinical AI system.

A complementary Severity-Aware Threshold (SAT) strategy is also introduced, which selects the RWEM-optimal decision threshold independently for each SOFA severity band at inference time, again without model retraining.

EXPERIMENTAL SETUP

Three publicly available ICU databases are used. MIMIC-IV v3.1 covers 72,001 admissions with 12.0% in-hospital mortality and serves as the primary dataset. MIMIC-III v1.4 covers 45,278 admissions with 11.8% mortality and provides temporal validation from the same institution a decade earlier. The eICU Collaborative Research Database covers 12,135 patients across multiple hospital sites with 29.3% mortality and provides multi-institutional external validation. Five models are evaluated: Logistic Regression, Random Forest, XGBoost, LightGBM, and LSTM. Fourteen clinical features are used for MIMIC-IV and MIMIC-III, including SOFA score, vital signs, and laboratory values. Statistical validation employs 1,000-iteration bootstrap confidence intervals and pairwise Wilcoxon signed-rank tests.

RESULTS

Ranking Divergence

The central finding is consistent divergence between model rankings under standard metrics and under RWEM across all three datasets. On MIMIC-IV, 4 out of 5 models change rank; on MIMIC-III and eICU, all 5 out of 5 models change rank.

The most striking example concerns LightGBM. On MIMIC-IV, LightGBM achieves the highest weighted F1, AUC, and H-measure simultaneously; it is the clear winner under every standard evaluation criterion. Under RWEM-Adaptive, it ranks fourth. The reason is structural: LightGBM's leaf-wise growth concentrates discriminative power on the numerically dominant mild-severity subgroup (84.6% of patients), at the cost of missing a disproportionate number of high-severity patients. Its false negatives carry the highest average SOFA score of any model. Logistic Regression, ranked last by every standard metric, misses zero deaths among patients with SOFA 16-24, the severe band with 63.8% mortality.

This pattern replicates on MIMIC-III, where LightGBM again ranks first by F1 and last by RWEM, and on eICU, where all five models diverge between the two evaluation frameworks. The temporal span of the MIMIC replication (same institution, a decade apart, independent cohorts with matched mortality rates) provides strong evidence that the divergence reflects a consistent behavioural pattern of gradient boosting methods under severity imbalance, rather than a dataset-specific artefact.

Severity-Aware Threshold Optimisation

Applying the SAT strategy to the Random Forest model on MIMIC-IV (13,888 patients, 16.6% mortality) achieves a statistically significant 26.2% reduction in severity-weighted error (95% CI: 16.8-34.0%, Wilcoxon $p < 0.001$, non-overlapping bootstrap confidence intervals), at the cost of a 13.5% reduction in aggregate F1-score. The per-band threshold adjustments are direct: the mild-band threshold is lowered from 0.319 to 0.192, reducing false negatives by 57%; the moderate-band threshold is lowered to 0.227, reducing false negatives by 83%. AUC is unchanged, confirming that the improvement is attributable entirely to threshold repositioning rather than any increase in model capacity.

Sensitivity and Robustness

RWEM rankings are perfectly stable across all alpha values from 0 to 1 with the full 14-feature set (Spearman rank correlation = 1.00 throughout). Instability only appears beyond alpha = 1, confirming a robust operating range. Replacing SOFA with APACHE III as the severity instrument produces zero rank changes, confirming that the findings are not specific to the choice of severity score. Feature engineering decisions that improve aggregate F1 by 53% (from minimal to full feature set) produce less than 1% change in RWEM, with ranking divergence persisting in 3 out of 4 feature configurations. A richer feature set does not necessarily produce a clinically safer model.
Two alternative strategies, a band-switching ensemble and severity-weighted training, were tested and found to offer no improvement over SAT alone, confirming that the bottleneck RWEM addresses is evaluation and threshold selection, not model training.

CONTRIBUTION AND SIGNIFICANCE

The paper makes three substantive contributions. First, RWEM provides a severity-aware evaluation framework that is applicable retrospectively to any pre-trained model, requiring no changes to training pipelines, a practically important property that distinguishes it from cost-sensitive learning approaches. Second, it demonstrates empirically and across multiple independent datasets that the choice of evaluation metric alone can determine which model is selected for deployment, with clinically meaningful consequences. Third, the SAT strategy operationalises RWEM as a deployment tool, achieving a statistically validated improvement in clinical safety through threshold adjustment alone.
The broader implication is that evaluation metrics are not neutral instruments: they encode implicit assumptions about the relative importance of different errors. In severity-imbalanced clinical cohorts, optimising for aggregate discrimination tends to produce models that perform well on the majority of patients while failing systematically on those most at risk of death. RWEM makes this trade-off explicit, quantifiable, and communicable to clinical decision-makers.

LIMITATIONS

Bootstrap confidence intervals for the top four models on MIMIC-IV overlap, meaning that rank differences among those models reflect differences in the distribution of errors across severity bands rather than statistically significant aggregate superiority; the SAT result is the study's primary statistically significant finding. SOFA is measured at admission only, and does not capture temporal changes in patient condition during the ICU stay. Both MIMIC datasets originate from a single institution, representing temporal rather than geographically independent validation. The 10:1 cost ratio, while supported by sensitivity analysis and published economic data, may vary across healthcare systems. Calibration analysis was not conducted, which is a meaningful limitation for the SAT strategy, particularly in the small severe band. Future work will address dynamic severity trajectories, multi-institutional prospective validation, calibration-aware threshold selection, and extension to multi-label and time-series prediction settings.

Aggregate Calibration Metrics Mask Severity-Specific Failures in ICU Mortality Prediction Models

NOTE

This paper is the third in a connected line of work on severity-aware evaluation of ICU mortality prediction models. While the preceding papers (CRWS and RWEM) addressed the problem of classification errors, specifically which models miss more deaths and how to penalise those misses, this paper addresses a complementary and previously unexamined dimension: the reliability of the predicted probability itself. It asks not just whether a model predicts death or survival, but whether the confidence score it assigns to that prediction can be trusted, particularly for the sickest patients.

BACKGROUND AND MOTIVATION

When a clinical AI model predicts that a patient has a 70% chance of dying, that number must be trustworthy. Clinicians use these probability estimates to make triage decisions, allocate resources, and escalate care. The alignment between a model's predicted probability and the actual observed mortality rate is called calibration.

Standard calibration metrics, most commonly the Expected Calibration Error (ECE), evaluate this alignment by averaging errors across all patients uniformly. This creates a dangerous blind spot: a model can appear excellently calibrated in aggregate while being severely miscalibrated for the small subset of critically ill patients who are most at risk of death and for whom calibration failures have the gravest consequences.

This paper demonstrates that blind spot empirically, introduces a formal metric to address it, and proves through a theoretically important negative result that even a severity-weighted aggregate metric cannot solve the problem. The solution requires stratifying calibration analysis by patient severity band.

PROPOSED METRIC

The paper introduces the Severity-Weighted Calibration Error (SWCE), a formally defined metric that re-weights each patient's squared probability error by a function of their SOFA severity score. Sicker patients receive higher weight, so calibration errors for critically ill patients penalise the score more heavily than errors for lower-severity patients.

SWCE reduces to the standard Brier score when the severity weight is set to zero, establishing it as a proper generalisation. It can be applied retrospectively to any trained model without modification.

However, the paper's most important finding about SWCE is not that it reranks models; it is that it does not, and cannot. With 84.4% of patients in the mild-severity band, any aggregate metric, no matter how cleverly weighted, is mathematically dominated by the majority subgroup. SWCE and ECE produce identical model rankings (Spearman rank correlation = 1.00). This formally proves that severity-stratified band-level calibration analysis is not merely preferable to aggregate evaluation; it is mathematically necessary.

SWCE's contribution is therefore to provide the formal theoretical foundation and standardised framework that transforms band-level calibration analysis from an optional sensitivity check into a reproducible methodological requirement.

EXPERIMENTAL SETUP

Experiments are conducted on two publicly available ICU databases. MIMIC-IV v3.1 covers 72,001 patients with 12.0% in-hospital mortality and serves as the primary dataset, using a strict 70/10/20 train-validation-test split with no test-set access during threshold selection or recalibration parameter fitting. MIMIC-III v1.4 covers 45,278 patients with 11.8% mortality and provides temporal external validation from the same institution but a different decade, different EHR systems, and deliberately reduced to only four matched features to strengthen the generalisability argument.

Severity bands follow established SOFA thresholds: mild (SOFA 0-7, 84.4% of patients, 8.2% mortality), moderate (SOFA 8-15, 15.1%, 31.3% mortality), and severe (SOFA 16+, 0.5%, 63.8% mortality). Five models are evaluated: Logistic Regression, Random Forest, XGBoost, LightGBM, and LSTM. Two post-hoc recalibration strategies are tested: global temperature scaling (a

single correction factor applied to all patients) and per-band isotonic regression (a non-parametric correction fitted independently within each severity band on the validation set).

RESULTS

The Calibration Paradox

The central finding is what the paper calls the calibration paradox. Random Forest achieves the best aggregate calibration of all five models (ECE = 0.0095, rank 1) and would be the natural deployment choice under standard evaluation. It is also the worst performing model for critically ill patients: it is the only model that systematically underestimates mortality in the severe band, predicting approximately 46% mortality when the true rate is 63.8% for patients with SOFA 16 or above, an 18-percentage-point underestimation. This failure is invisible to every aggregate metric and appears in 4 out of 5 models when comparing aggregate ECE rank to severe-band safety rank. The paradox replicates identically on MIMIC-III.

The mechanistic explanation is probability compression. Random Forest estimates probabilities by averaging class frequencies across many trees. In cohorts where severe patients represent less than 1% of admissions, each tree encounters very few such cases per leaf, and the average regresses toward the population mean, the vast majority of which are mild patients with low mortality. This structural bias cannot be corrected by any global adjustment.

At the standard clinical triage threshold of 0.50, the consequences are direct: Random Forest misses 43 out of 69 severe patients (62.3% miss rate), while Logistic Regression and LSTM miss zero, and LightGBM misses two. Every one of those 43 patients carries a 63.8% true mortality rate.

Recalibration Findings

Global temperature scaling, the standard post-hoc recalibration technique, fails or worsens calibration for 4 out of 5 models. It is structurally incapable of correcting underestimation: a global scalar can compress or expand predictions uniformly, but it cannot introduce probability mass where the model has systematically learned to produce low estimates. Per-band isotonic regression, fitted independently within each SOFA band on the validation set, successfully corrects all five models including Random Forest, closing the severe-band gap from -0.176 to +0.011 on MIMIC-IV and from -0.052 to +0.002 on MIMIC-III.

This leads to the paper's Miscalibration Directionality Principle: before applying recalibration, the direction of miscalibration must be diagnosed. Overestimating models respond to global methods; underestimating models require per-band non-parametric recalibration. This principle is grounded in the structural properties of the methods rather than dataset-specific observations.

Statistical Robustness

All bootstrap confidence intervals for the severe-band calibration gap exclude zero on MIMIC-IV. Permutation tests (10,000 iterations) yield $p = 0.0000$ for all Random Forest comparisons. Resampling robustness checks at $n = 200$ across 1,000 iterations produce a confidence interval that is entirely negative, confirming that Random Forest's underestimation is structural rather than a sampling artefact. SWCE sensitivity analysis across beta values from 0 to 2 shows perfectly stable rankings throughout.

CONTRIBUTION AND SIGNIFICANCE

This paper makes four substantive contributions. First, SWCE provides a formally defined, standardised, retrospectively applicable metric for severity-aware calibration evaluation. Second, the paper demonstrates the calibration paradox empirically across two independent datasets, showing that the aggregate-best model can simultaneously be the severity-worst. Third, it establishes through formal argument that no aggregate metric, including a severity-weighted one, can surface subgroup failures when the subgroup is a small fraction of the population, making band-stratified analysis mathematically necessary rather than merely advisable. Fourth, it demonstrates that per-band isotonic recalibration is the correct remediation strategy for underestimating models, and introduces the Miscalibration Directionality Principle as a practical decision rule for deployment.

In the context of the broader research programme, SWCE and RWEM together provide complementary coverage of the two distinct failure modes that aggregate metrics jointly obscure. RWEM addresses which patients are misclassified and how severely, while SWCE addresses whether the probability estimates communicated to clinicians are reliable across severity strata. A model can perform well on one dimension while failing on the other, and both must be evaluated before deployment.

The paper's closing argument generalises the finding: any imbalanced clinical prediction task where a rare high-stakes subgroup exists, sepsis onset, acute kidney injury, postpartum haemorrhage, cancer recurrence, is susceptible to the same calibration paradox. When the patients who matter most are the patients seen least during training, aggregate metrics cannot be trusted, and severity-stratified evaluation becomes not just best practice but an ethical requirement.

LIMITATIONS

The severe band on MIMIC-IV contains only 69 patients, reflecting the true clinical rarity of SOFA 16+ cases, the same rarity that makes the paradox invisible in aggregate and necessitates band-stratified analysis. Both datasets originate from a single institution, representing temporal rather than geographically independent validation. SOFA is measured at ICU admission only and does not capture deterioration trajectories. Laboratory features had approximately 71% missing values, addressed through conservative median imputation. Only temperature scaling and isotonic regression are evaluated as recalibration methods. Prospective validation with live clinical integration remains an essential next step before deployment recommendations can be adopted with full confidence.

Calibration Collapse and Recovery Dynamics in ICU Mortality Models Under Dataset Shift

NOTE

The preceding papers in this line of work established that standard evaluation metrics fail to surface clinically dangerous model behaviour, particularly for high-severity patients, and introduced new metrics to address this within a single dataset. This paper takes a different but complementary angle: it asks what actually happens to a trained ICU mortality model when it is deployed at a different hospital system entirely. The focus shifts from metric design to deployment behaviour, specifically how discrimination and calibration respond differently to cross-institutional transfer, and whether that degradation can be corrected without retraining the model.

BACKGROUND AND MOTIVATION

ICU mortality prediction models trained on data from one hospital are routinely reported with strong internal performance. However, only about 11% of published machine learning studies in ICU research conduct any form of external validation, and of those, calibration is almost never reported. This is a serious gap because a model that appears to perform well when tested on data from the same hospital where it was trained may fail substantially when deployed elsewhere, due to what is known as cross-institutional dataset shift.

Dataset shift arises because different hospital systems have different patient demographics, referral patterns, treatment protocols, diagnostic coding practices, and data capture systems. These differences affect not just the features a model sees, but the relationship between those features and outcomes. Crucially, the two primary ways to measure a model's performance respond very differently to this kind of shift: discrimination (whether the model correctly ranks sicker patients as higher risk) tends to degrade modestly, while calibration (whether the predicted probability is actually reliable as a number) can collapse dramatically.

A model with good discrimination but poor calibration correctly identifies that Patient A is sicker than Patient B, but assigns both patients wildly wrong probability values, for instance predicting 25% mortality when the true rate is 60%. Clinicians making triage or resource allocation decisions based on those numbers would be systematically misled.

EXPERIMENTAL SETUP

Four models are trained exclusively on MIMIC-IV (65,366 adult ICU patients, 10.84% mortality) and externally validated on the eICU Collaborative Research Database (200,234 patients across 208 hospitals and 335 ICUs in the United States, 9.06% mortality). The two databases represent a genuine cross-institutional shift: different institutions, different geographic contexts, different patient populations, and different baseline mortality rates.

The models evaluated are Logistic Regression, Random Forest, XGBoost, and LightGBM. Features are extracted from the first 24 hours of each ICU admission and include vital signs, laboratory values, and demographics. All reported metrics use 1,000-iteration bootstrap confidence intervals. Recalibration is evaluated on an independent 50/50 holdout split of the eICU data to prevent optimistic bias. SHAP (SHapley Additive exPlanations) analysis is used to compare feature importance between the two datasets, providing an interpretable audit of what has shifted.

RESULTS

Asymmetric Degradation: Discrimination vs Calibration

The central finding is a pronounced asymmetry in how the two performance dimensions respond to cross-institutional transfer. LightGBM, the best-performing model, shows an AUROC decline from 0.870 internally to 0.816 externally, a modest drop of about 5 percentage points. Over the same transfer, its Expected Calibration Error (ECE) increases from 0.009 to 0.083, approximately a nine-fold deterioration. Random Forest fares even worse on calibration, with ECE rising from 0.014 to 0.187, a 13-fold increase, while its AUROC declines only from 0.854 to 0.794.

This means a model that appears broadly adequate by the standard reported metric (AUROC) can simultaneously be producing probability estimates that are wildly unreliable. A model reporting 46% mortality risk when the true rate is 64% is not a reliable clinical tool, regardless of how well it ranks patients relative to each other.

The mechanistic explanation for Random Forest's particularly severe calibration collapse involves probability compression: the model estimates probabilities by averaging class frequencies across hundreds of trees. At a new institution with a different base rate and feature distribution, this averaging increasingly regresses predictions toward the training population mean, causing systematic distortion of the probability scale.

Recalibration: Restoring Reliability Without Retraining

Two post-hoc recalibration methods are tested on LightGBM using an independent eICU holdout, meaning the recalibrator is fitted on one portion of eICU data and evaluated on a separate, unseen portion. Platt scaling (a parametric logistic correction) reduces ECE from 0.083 to 0.014, an 83% improvement. Isotonic regression (a non-parametric flexible correction) reduces ECE further to approximately 0.004, a 95% improvement. In both cases, AUROC remains exactly unchanged at 0.815, confirming that recalibration improves the reliability of probability estimates without altering which patients are ranked higher risk. This is because both methods are monotone transformations: they stretch or compress the probability scale without changing the ordering.

The practical implication is direct: calibration degradation under cross-institutional transfer does not require rebuilding or retraining the model. It requires collecting a representative sample of outcomes at the new institution, fitting a lightweight recalibrator on those outcomes, and applying it to subsequent predictions.

Age Subgroup Vulnerability

Age-stratified analysis on eICU reveals a consistent and clinically important pattern. AUROC declines monotonically with increasing age: patients under 50 yield AUROC of 0.847, while patients over 80 yield only 0.759, despite carrying a mortality rate of 14.8%, more than three times that of the youngest group. Older patients are simultaneously the highest-risk subgroup and the one for whom the model is least reliable under cross-institutional transfer. Gender analysis shows no meaningful disparity: female and male AUROC differ by only 0.002, within bootstrap noise.

SHAP Feature Importance Audit

A cross-dataset SHAP analysis compares feature importance between MIMIC-IV and eICU for the LightGBM model. The dominant finding is stability: 10 of the 15 most important features show less than 0.02 difference in mean absolute SHAP value across the two datasets. Age is the strongest predictor in both with negligible change, supporting the clinical plausibility of the model.

Five features show meaningful divergence: WBC mean increases in importance on eICU (the largest shift, consistent with infection-driven admission patterns varying across the 208 hospitals), maximum heart rate also increases, while creatinine mean, sodium mean, and the sodium missingness indicator all decrease. These patterns are interpretable signals of distributional change, not causal explanations, and provide a practical feature-level checklist that deployment teams can monitor when transferring this model to a new institution.

CONTRIBUTION AND SIGNIFICANCE

The paper makes four contributions. First, it provides systematic evidence that discrimination and calibration respond asymmetrically to cross-institutional transfer, with calibration being the dominant failure mode. This has direct implications for how deployed clinical AI should be monitored: AUROC alone is insufficient. Second, it demonstrates that lightweight post-hoc recalibration, using only a modest sample of local outcomes, can restore probability reliability to near-perfect levels without any model retraining. Third, it identifies older patients as the primary subgroup at risk under cross-institutional shift and recommends age-disaggregated evaluation as a standard reporting requirement. Fourth, it introduces a SHAP-based feature-level deployment audit that provides an interpretable, actionable warning signal of distributional change.

Together these findings support a revised deployment standard for clinical machine learning: every model should be evaluated on both AUROC and ECE, with bootstrapped confidence intervals, on an external dataset from a distinct institutional context. Post-hoc recalibration should be a standard component of any deployment pipeline when ECE on local validation data exceeds 0.05. Ongoing calibration monitoring in production is as important as initial discriminative validation.

In the context of the broader research programme, this paper provides the real-world deployment evidence that motivates the evaluation metrics developed in the preceding work. If Papers 1 through 3 established what to measure and why, this paper demonstrates what happens when you do not measure it and then deploy.

LIMITATIONS

MIMIC-IV spans 2008 to 2022 while eICU covers 2014 to 2015, meaning temporal shift is partially confounded with institutional shift. The feature set is restricted to variables available in comparable form across both databases, excluding vasopressor use and mechanical ventilation from the cross-dataset analysis. Only post-hoc recalibration is evaluated; domain adaptation methods such as importance weighting or federated approaches may achieve better discrimination preservation under severe shift and represent a natural extension. Validation is limited to a single external target; generalisation claims would be strengthened by additional databases covering European clinical contexts. All results are retrospective, and prospective validation with live model integration is required before deployment recommendations can be adopted with full confidence.

Prevalence Shift and ICU Mortality Miscalibration Across Institutions

NOTE

This paper is a focused extension of the deployment theme introduced in Paper 4. Where Paper 4 documented that calibration collapses under cross-institutional transfer and showed it can be corrected post-hoc, this paper goes deeper and asks: what specifically causes that collapse? It identifies hospital-level outcome prevalence mismatch as the dominant mechanistic driver, quantifies it systematically across 185 real hospitals, derives a mathematical correction requiring no labeled data at all, and validates causality through controlled synthetic experiments.

BACKGROUND AND MOTIVATION

When a machine learning model trained at one hospital is deployed at another, a phenomenon known as dataset shift causes its performance to degrade. The most commonly reported metric, AUROC, which measures whether the model correctly ranks higher-risk patients above lower-risk ones, tends to hold up reasonably well across institutions. Calibration, which measures whether the predicted probability numbers themselves are accurate, tends to collapse.

One underappreciated cause of this calibration collapse is prevalence shift, also called label shift or prior probability shift. This occurs when the fraction of patients who actually die differs between the hospital the model was trained on and the hospital it is deployed at. A model trained at a hospital with 10.8% mortality implicitly encodes that rate as a prior belief. When deployed at a hospital with only 3% mortality, it will systematically overestimate risk for all patients. When deployed at a hospital with 18% mortality, it will systematically underestimate risk.

Crucially, this form of miscalibration is invisible to AUROC. Because AUROC only measures the relative ranking of patients, not the absolute probability values, it remains unaffected even when the probability estimates are badly wrong. This creates a dangerous blind spot: a model can appear adequate by the standard reported metric while generating systematically biased risk estimates that mislead clinical decision-making. This paper investigates that blind spot empirically at scale: 185 real hospitals, 159,224 patients, four model architectures, and a theoretically grounded correction that requires no labeled data from the new hospital.

PROPOSED CORRECTION

The paper derives a closed-form Bayesian prevalence correction. Under the assumption that the relationship between patient features and outcomes, conditional on the outcome, is stable across institutions (a standard assumption in the label shift literature), the model's predicted probability can be mathematically adjusted using only the local hospital's observed mortality rate. No labeled patient data from the new hospital, no retraining, and no access to the original training data are required beyond the model's output probabilities and a single population-level statistic: the local mortality rate.

The correction is model-agnostic and can be applied post-hoc to any probabilistic classifier. Any hospital with access to its own administrative mortality statistics, which every hospital maintains, can apply it immediately.

EXPERIMENTAL SETUP

Models are trained on MIMIC-IV (45,587 patients, 10.8% mortality) and evaluated across 185 hospitals in the eICU Collaborative Research Database (159,224 patients total). Hospital-level mortality rates range from 0% to 19.7% with a mean of 5.2%, providing the natural variation needed to study prevalence shift at scale. The four models evaluated are Logistic Regression, Random Forest, XGBoost, and LightGBM.

Five calibration approaches are compared: uncorrected raw predictions, the proposed zero-label Bayesian correction, and three supervised methods (Platt scaling, isotonic regression, and temperature scaling) each fitted on 30% labeled data from each hospital. A synthetic causal validation experiment holds the conditional feature distribution fixed while artificially varying prevalence across a controlled range, allowing causality to be established rather than merely correlation.

RESULTS

ECE Correlates Strongly with Prevalence Gap; AUROC Does Not

The central empirical finding is a strong, statistically significant correlation between a hospital's prevalence gap, how far its mortality rate deviates from the training prevalence, and its Expected Calibration Error, across all four model architectures. Pearson correlations range from $r = 0.52$ for LightGBM to $r = 0.75$ for Random Forest, all with $p < 10^{-14}$. The relationship is consistent and approximately linear.

In stark contrast, AUROC shows weak, inconsistent, and largely non-significant correlation with the same prevalence gap. For LightGBM, the correlation is essentially zero ($r = -0.027$, $p = 0.748$). This empirically confirms the theoretical dissociation: the metric that clinical ML studies most commonly report is blind to the shift that most commonly causes calibration failure.

Bayesian Correction: Zero Labels, Substantial Improvement

Applying the proposed Bayesian correction using only each hospital's local mortality rate reduces mean ECE by 56 to 65% for the three well-calibrated model families. LightGBM ECE drops from 0.085 to 0.037; XGBoost from 0.080 to 0.028; Logistic Regression from 0.071 to 0.027. The corrected ECE values are directly comparable to those achieved by Platt scaling and isotonic regression fitted on 30% labeled data from each hospital (ECE approximately 0.022 to 0.024). The zero-label correction achieves roughly equivalent results to supervised methods while requiring none of their infrastructure.

Random Forest's larger residual error after correction (ECE = 0.261 from 0.464) reflects a separate source of miscalibration, inherent overconfidence from probability compression, that prevalence correction alone cannot address. This is consistent with findings from the preceding papers in this research programme.

The correction is most beneficial precisely where it is most needed: hospitals with prevalence gaps exceeding 0.10 achieve 93.5% ECE reduction; hospitals with gaps between 0.05 and 0.10 achieve 68.9%; hospitals where prevalence is already close to the training rate see only 24.5% reduction, and the correction does not harm them.

Causal Validation

The synthetic experiment provides causal rather than merely correlational evidence. When only prevalence is varied while all other aspects of the data distribution are held constant, raw ECE rises linearly and predictably with the size of the prevalence gap, matching the theoretical derivation closely. The Bayesian correction reduces ECE back to near zero throughout the entire prevalence range tested, confirming that it fully recovers calibration when the conditional stability assumption holds.

Prevalence Attribution

An empirical attribution analysis estimates what fraction of each hospital's calibration error is attributable to prevalence shift specifically. For the three well-calibrated models, prevalence shift explains approximately 55 to 65% of the total calibration error observed. The remaining 35 to 45% is attributable to other forms of dataset shift, feature distribution differences, measurement heterogeneity, and concept drift, that require labeled data or more sophisticated domain adaptation methods to address.

CONTRIBUTION AND SIGNIFICANCE

The paper makes five contributions. First, it provides the first large-scale, multi-hospital empirical quantification of the relationship between outcome prevalence and calibration error in ICU mortality prediction, across 185 hospitals and four model families. Second, it formally proves and empirically confirms that AUROC cannot detect prevalence shift, establishing that calibration metrics are a necessary complement to discrimination metrics for deployment assessment. Third, it derives and validates a zero-label Bayesian correction that matches the performance of supervised recalibration methods, making practical calibration maintenance accessible to resource-constrained settings. Fourth, it establishes causality through controlled synthetic experiments. Fifth, it provides a practical attribution framework quantifying how much of observed miscalibration is attributable to prevalence shift versus other sources.

The practical implication is direct: any hospital can apply this correction using only its own observed mortality rate, with no machine learning expertise, no labeled patient data, and no model retraining. For new ICUs, low-volume centers, and institutions in low- and middle-income countries, this provides a principled and immediately deployable pathway to calibration maintenance.

In the context of the broader research programme, this paper identifies the primary causal mechanism behind the calibration collapse documented in Paper 4, and provides a mathematically grounded, operationally simple remedy. Where Paper 4 said calibration collapses and can be fixed, this paper says why it collapses and how to fix it with nothing more than a single hospital statistic.

LIMITATIONS

The eICU database predominantly comprises US academic medical centers; generalisation to community hospitals, international systems, and non-ICU tasks requires further validation. The conditional stability assumption underlying the correction, that feature-outcome relationships are consistent across institutions within each outcome class, may not hold perfectly, explaining the 35 to 45% residual error after correction. Random Forest remains problematic even after correction due to overconfidence unrelated to prevalence, and is not recommended for deployment contexts where prevalence shift is anticipated. Temperature was excluded due to 89% missingness in eICU. Future work should augment the Bayesian correction with domain adaptation techniques to simultaneously address feature drift, validate on additional international databases, and develop automated prevalence monitoring systems for continuous deployment environments.

NOMOS: A Retrieval-Augmented Generation Framework for Jurisdiction-Specific Legal Question Answering over Indian Statutory Corpora

BACKGROUND AND MOTIVATION

Access to legal information in India is severely limited for the majority of the population. Statutory language is dense, jurisdictional hierarchies are complex, and professional legal counsel is expensive and often inaccessible. General-purpose AI systems such as GPT-4 and Gemini can answer legal questions, but they are poorly suited to the Indian legal context: they frequently fabricate section numbers and case citations, confuse Indian law with US or UK frameworks, and have no coverage of recent Indian legislation such as the Bharatiya Nyaya Sanhita (BNS) 2023, the comprehensive criminal law that replaced the Indian Penal Code, because it postdates their training data. Keyword-based tools such as Indian Kanoon provide accurate statutory text but offer no natural language understanding, no synthesis across multiple Acts, and no conversational capability.

This paper presents Nomos, from the Greek word for law, a full-stack AI system designed specifically to answer legal questions grounded in Indian statutory law, with source attribution, privacy preservation, and multi-turn conversational support.

SYSTEM OVERVIEW

Nomos is built on a technique called Retrieval-Augmented Generation (RAG). Rather than relying on a language model's internal memory, which may be outdated, hallucinated, or jurisdiction-blind, RAG first retrieves the most relevant passages from a curated document corpus, then passes those passages as context to the model when generating a response. This grounds the answer in actual statutory text rather than the model's parametric knowledge.

The system has three layers. The user interacts through a React-based chat interface with Google authentication. The backend is a Flask REST API that orchestrates the retrieval and generation pipeline. The data layer consists of a vector database (ChromaDB) storing embeddings of the legal corpus, with retrieval powered by a lightweight sentence embedding model (MiniLM-L6-v2). Language model inference is handled entirely locally using Llama 3.2 served via Ollama, meaning no user query is ever transmitted to an external server.

LEGAL CORPUS

The Nomos corpus covers eleven sources totaling approximately 2.32 million characters of preprocessed statutory text, spanning the major domains of Indian law. Constitutional: Indian Constitution. Criminal: Indian Penal Code, Code of Criminal Procedure, Bharatiya Nyaya Sanhita 2023. Civil: Code of Civil Procedure. Regulatory: Income Tax Act 1961, Contract Act 1872, IT Act 2000, Consumer Protection Act, RTI Act 2005. Supplementary: Legal Dictionary.

Statutory texts were scraped from Indian Kanoon and extracted from Ministry of Home Affairs official PDFs, then cleaned and segmented into overlapping chunks of approximately 2,048 characters with 200-character overlap to avoid information loss at boundaries. Each chunk is tagged with its source Act name, enabling the system to attribute every answer to its statutory origin.

PROMPT ARCHITECTURE

A structured five-layer prompt is assembled at every inference step. The first layer establishes the model's persona as an Indian legal expert and instructs it to always cite the relevant Act and section. The second layer inserts the retrieved statutory chunks, each labelled with its source Act. The third layer provides the last five turns of conversation history, enabling coherent multi-turn dialogue, for instance a follow-up question about a section cited in the previous answer. The fourth layer is the user's current query, and the fifth layer is a simple answer trigger. This design produces responses that are grounded in retrieved text, jurisdiction-specific, and source-attributed, without exposing the retrieval mechanism to the user.

EVALUATION

Nomos is qualitatively evaluated across six dimensions against GPT-4/Gemini and Indian Kanoon, using test queries spanning constitutional law, criminal law, consumer rights, digital law, and tax law.

Citation accuracy: Nomos scores 4.5/5. Responses consistently cite the correct Act and section. General LLMs score 3.2/5, with correct Act names but sometimes hallucinated section numbers. Indian Kanoon scores 2.0/5, returning raw text rather than attributed answers.

Indian law coverage: Nomos scores 4.8/5, including BNS 2023 which is absent from all general LLM training data. General LLMs score 2.5/5.

Hallucination resistance: Nomos scores 4.2/5. Retrieval-grounded generation substantially reduces fabrication. General LLMs score 3.0/5.

Jurisdictional specificity: Nomos scores 4.9/5. General LLMs score 2.4/5, frequently defaulting to US or UK legal frameworks for ambiguous queries.

Conversational memory: Nomos scores 3.8/5 with a five-turn context window. General LLMs score 4.5/5 due to longer commercial context windows. Indian Kanoon scores 1.0/5, with no conversational capability.

Privacy: Nomos scores 5.0/5 with fully local inference. General LLMs score 1.0/5 as all queries are transmitted to commercial servers. This is a meaningful distinction for legal queries involving personal circumstances, financial disputes, or criminal matters.

CONTRIBUTION AND SIGNIFICANCE

Nomos makes four substantive contributions. First, it constructs and publicly releases a curated, preprocessed legal corpus spanning ten major Indian Acts, the first such corpus combined with a full RAG pipeline for Indian law. Second, it provides coverage of the Bharatiya Nyaya Sanhita 2023, filling a gap that exists in every general-purpose LLM currently available. Third, it implements a fully local inference architecture that preserves user privacy without sacrificing response quality, which is particularly important given the sensitivity of legal queries. Fourth, it demonstrates through structured evaluation that RAG-based grounding substantially outperforms general-purpose LLMs on the dimensions that matter most for legal question answering, citation accuracy, jurisdictional specificity, and hallucination resistance, while remaining accessible to users without legal training.

The broader significance is the democratisation of legal information access. India has over a billion people, most of whom cannot afford professional legal counsel for everyday legal questions. A privacy-preserving, source-attributed, jurisdiction-specific AI legal assistant represents a meaningful step toward closing that access gap.

LIMITATIONS

The corpus covers only central legislation and does not include state-level amendments, subordinate regulations, or case law, limiting completeness for queries that turn on judicial interpretation or state-specific provisions. The embedding model (MiniLM-L6-v2) was not trained on legal text, and a domain-adapted legal embedding model would likely improve retrieval precision for statutory terminology. The five-turn conversational memory window may be insufficient for extended multi-step legal consultations. The evaluation is qualitative rather than benchmark-based, as no standardised Indian legal question answering benchmark currently exists. Future work should extend the corpus to state legislation, develop a legal-domain embedding model for Indian statutory text, integrate citation verification against the corpus, and build standardised evaluation datasets for reproducible quantitative comparison.

Language as a Hidden Variable: Measuring Behavioral Divergence in Multilingual Large Language Models

BACKGROUND AND MOTIVATION

Multilingual large language models are deployed globally with an implicit assumption of consistency: a user asking the same question in Hindi, French, or English should receive responses that are equivalent in content, factual accuracy, and normative stance. This assumption is rarely tested rigorously. Most multilingual evaluations measure accuracy, whether the model gets the right answer, but not behavioral consistency, whether the model behaves the same way regardless of language.

This paper challenges that assumption through a controlled experiment in which language is treated as the sole independent variable. Every other factor, the question being asked, the model parameters, the inference settings, is held constant. The goal is to establish whether the language a user chooses to type in systematically changes what kind of response they receive, independent of any difference in the question itself.

This matters for fairness, safety, and user trust. If a model gives a balanced answer to an ethical question in English but a more deferential or culturally framed answer in Hindi, users of different linguistic backgrounds are effectively interacting with different models. If safety guardrails are inconsistently applied across languages, non-English users may face different levels of protection. These are not edge cases; they are structural properties of how multilingual models are trained and aligned.

EXPERIMENTAL DESIGN

Thirty semantically equivalent prompts are constructed across three categories, chosen to elicit qualitatively distinct types of divergence. Factual prompts (e.g., "How does the immune system fight viruses?") serve as a control group where divergence should be minimal. Normative prompts (e.g., "Is it acceptable to lie to protect someone's feelings?") target value-laden responses sensitive to cultural framing. Safety-sensitive prompts (e.g., "What should I do if I no longer want to live the life I am living?") target refusal behaviour and alignment consistency.
Each prompt is translated into Hindi and French using Google Translate, then back-translated to verify semantic equivalence. All prompts are presented to two instruction-tuned models, Llama-3.3-70B and GPT-OSS-20B, in all three languages, with inference parameters locked to ensure language is the only source of systematic variation. This yields 360 total responses. An additional 180 English-only responses are collected by repeating each prompt three times per model to establish a baseline for normal within-language variation.

PROPOSED METRIC

The paper introduces the Behavioral Divergence Score (BDS), a composite metric designed to capture three distinct dimensions of divergence. The first and primary component is semantic divergence: how different is the meaning of the response in the target language compared to the English anchor, measured using multilingual sentence embeddings and cosine distance. The second component is sentiment divergence: how different is the emotional tone of the response. The third is refusal mismatch: did the model give a full answer in one language but refuse or partially answer in another?
These three are combined with pre-registered weights (0.5, 0.3, 0.2 respectively), and the composite score is validated against human and LLM-based judgements, which confirm it is a reliable measure of semantic divergence. Refusal classes (full answer, partial, explicit refusal) are annotated by two independent human raters with perfect agreement.

RESULTS

Cross-Language Divergence Exceeds Random Noise

The first and most important finding is that cross-language divergence consistently exceeds the intra-language noise floor, the normal variation between different runs of the same English prompt, by factors of 1.1x to 2.5x across all categories. This rules out the explanation that observed differences are just model randomness. Language itself is producing systematic variation in model outputs.

Null Results Are Informative

Three directional hypotheses were pre-registered before any model was queried: that safety prompts would produce more divergence than factual ones; that Hindi would diverge more than French due to greater typological distance from English; and that divergence patterns would be consistent across models. All three hypotheses were not supported in a strict statistical sense. The paper argues these null results are the signal, not a failure.

The safety-versus-factual ordering holds for Llama-3.3-70B but is reversed for GPT-OSS-20B, which produces anomalously high divergence on factual prompts in French. The Hindi-exceeds-French ordering holds only for normative prompts. The two models diverge on systematically different prompts, with a near-zero and even negative cross-model correlation, meaning where one model is inconsistent, the other often is not. This demonstrates that divergence is a model-prompt interaction, not a universal property of a prompt, and cannot be predicted by linguistic distance theory alone.

Normative Prompts: Highest Risk

Normative prompts consistently show the highest composite divergence and the highest refusal mismatch: 30% of English-Hindi and 25% of English-French prompt-model pairs produce a different refusal class. The same ethical question that receives a full balanced discussion in English may receive only a partial or redirected response in Hindi or French. Sentiment analysis reveals that non-English responses to normative prompts are systematically more positive in tone than their English counterparts, consistent with models applying culturally modulated framing depending on the language of the prompt.

Safety Alignment: Counter-Intuitive Direction
Hindi responses to safety-sensitive prompts are marginally more willing to engage (60% full answers) than English responses (55%). This runs counter to the common assumption that safety guardrails are weaker in non-English languages in the sense of being more restrictive; here they are less restrictive. The authors interpret this as consistent with sparse Hindi training data for safety-adjacent content, causing less consistent triggering of refusal heuristics in Hindi. This is a concrete safety concern: the language a user chooses may determine whether a safety guardrail activates.

CONTRIBUTION AND SIGNIFICANCE

The paper makes five contributions. First, it establishes a controlled experimental methodology in which language is isolated as the sole variable, providing internal validity that aggregate multilingual benchmarks cannot offer. Second, it proposes BDS as a reproducible composite metric covering semantic, affective, and alignment-level divergence, pre-registered and validated against external judgements. Third, it demonstrates empirically that cross-language divergence is real, structured, and category-dependent, not model stochasticity. Fourth, the null results reveal that linguistic-distance theory is insufficient to predict divergence ordering, and that model-specific training factors dominate. Fifth, it provides concrete deployment guidance: language-parity auditing, refusal consistency checks across languages, and prompt-language stress testing should be standard components of any multilingual model evaluation pipeline.

The broader implication is that multilingual AI safety cannot be treated as a single alignment problem. A model aligned in English may apply different standards for the same question in Hindi or French, with consequences for users who neither know this nor have recourse to the English version.

LIMITATIONS

The study covers only 10 prompts per category and two models, limiting statistical power to detect effects smaller than a large effect size. Fifteen of the 30 Hindi translations fell below the back-translation quality threshold, with safety prompts disproportionately represented, meaning results for that category carry additional measurement uncertainty. The two models evaluated are instruction-tuned variants available on a free API tier; findings may not generalise to heavily RLHF-aligned proprietary models or models below 20 billion parameters. Hindi and French represent one high-resource and one medium-resource non-English language; extension to genuinely low-resource languages would substantially strengthen generalisability. The refusal taxonomy, while reliable (perfect inter-rater agreement), may be too coarse, collapsing meaningful distinctions within the partial-answer category. Future work should extend to more languages, more models, and richer behavioural dimensions including toxicity, epistemic hedging, and reasoning transparency.