# Environment-Conditional Interpretation Consistency

## Stabilizing LLM Explanations in Autonomous Driving Under Diverse Conditions

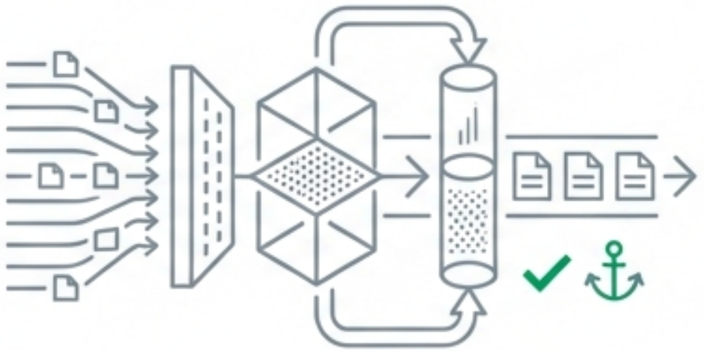
BASED ON THE ECIC FRAMEWORK & KDD ‘26 PROCEEDINGS

# Executive Summary: The Interpretability Gap

## The Challenge

Frontier LLMs achieve near-perfect driving decisions but suffer from reasoning instability.
Environmental shifts like rain or fog cause models to hallucinate different explanations for identical safety risks, breaking the "Interpretability Contract" required for trust.

## The Solution



The ECIC Framework disentangles decision-relevant features from environmental context. By utilizing a composite Consistency Index (CI) and "Contrastive Explanation Anchoring", we stabilize model reasoning without retraining the core foundation model.

## The Result

**93%**

Reduction in Faithfulness Gap

- 100% Pass Rate on structural consistency checks.
- Identification of critical Phase Transitions (visibility < 200m).
- Verified across 450 pairwise environmental conditions.

# Perfect Decisions, Unstable Reasons

## The Interpretability Contract Breach

```
DECISION: BRAKE (1.0)
EXPL: "Braking for pedestrian
at 25m."
```

Inconsistent Reasoning →

```
DECISION: BRAKE (1.0)
EXPL: "Braking due to wet road
surface reducing friction."
```

**Interpretability Consistency:** The stability of explanations when the decision scenario is constant ($s$), but environmental parameters ($e$) vary.

# Parameterizing the Environment

$$c_e = (v, p, l, f) \in \mathbb{R}^4$$

$v$ Visibility

10m (Fog) — 1000m (Clear)

$p$ Precipitation

0.0 (Dry) — 1.0 (Blizzard)

$l$ Light

0.0 (Night) — 1.0 (Day)

$f$ Friction

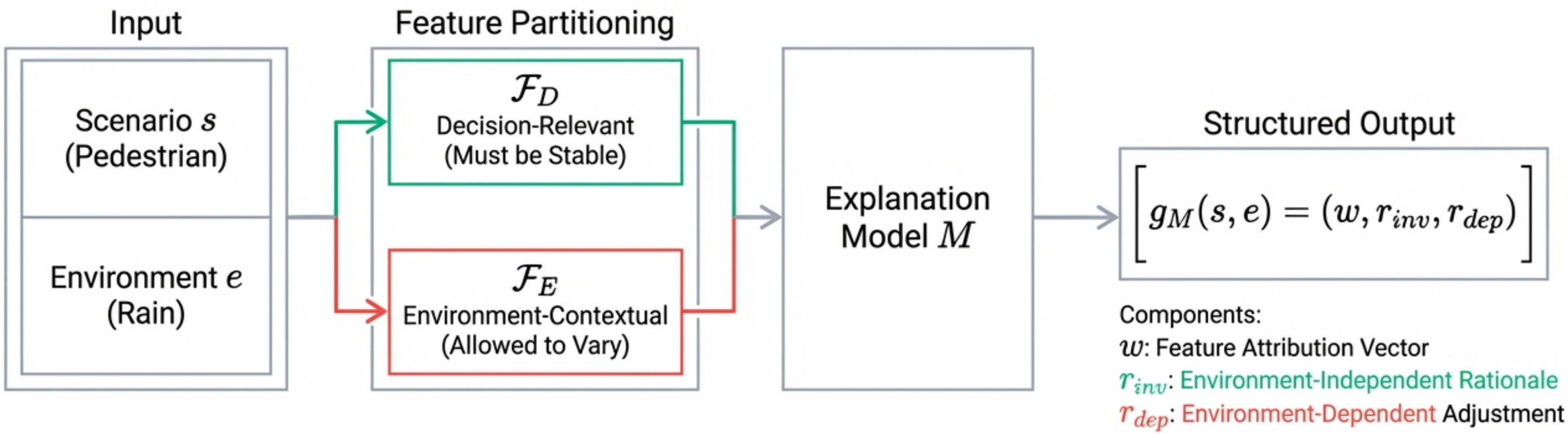
0.0 (Ice) — 1.0 (Asphalt)

Severity Formula

$$sev(e) = 1 - \frac{1}{4}\left(\frac{v}{1000} + (1 - p) + l + f\right)$$

Calculates environmental severity from 0 (**Benign**) to 1 (Extreme).

# The ECIC Framework Architecture

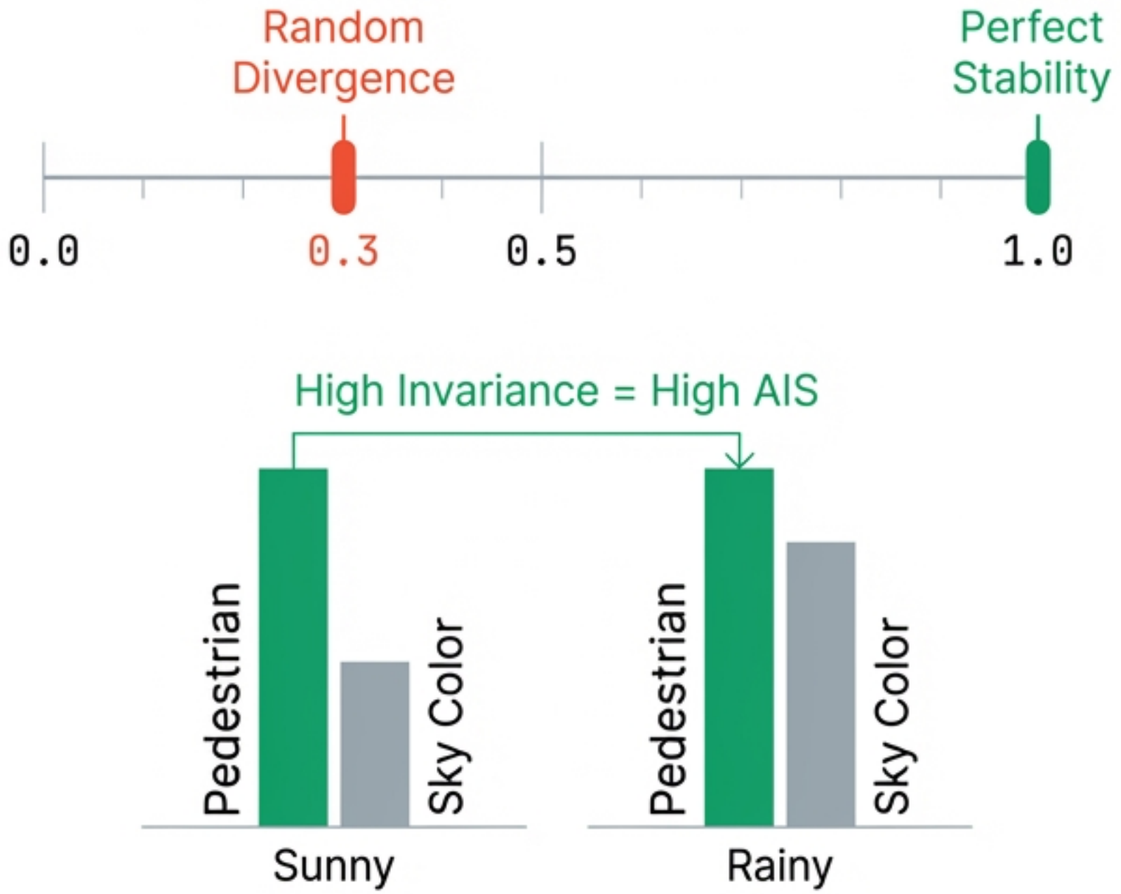


# Metric 1: Attribution Invariance Score (AIS)

Does the model “look” at the same features regardless of weather?

$$AIS(e_1, e_2 \mid s) = 1 - \mathrm{JSD}\big(w_D(s, e_1) \parallel w_D(s, e_2)\big)$$

Uses Jensen-Shannon Divergence (JSD) to measure the stability of feature weights. JSD is chosen for its symmetry and boundedness.

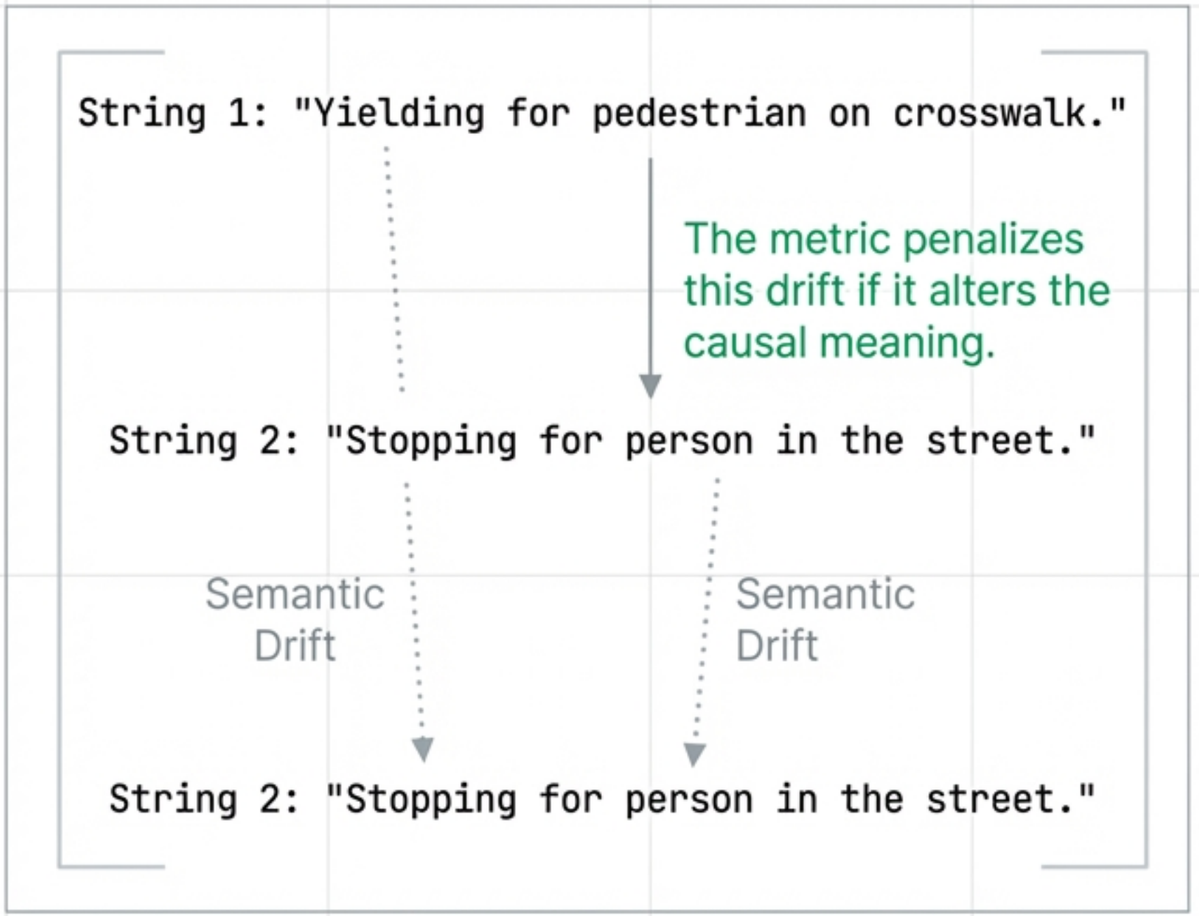


# Metric 2: Explanation Semantic Similarity (ESS)

Verifying the consistency of the language layer.

$$ESS(e_1, e_2 \mid s) = sim(r_{inv}(s, e_1), r_{inv}(s, e_2))$$

Measures textual consistency of the invariant rationale ($r_{inv}$) using Token-Level Jaccard Similarity. This ensures the LLM does not hallucinate different justifications for the same core decision.

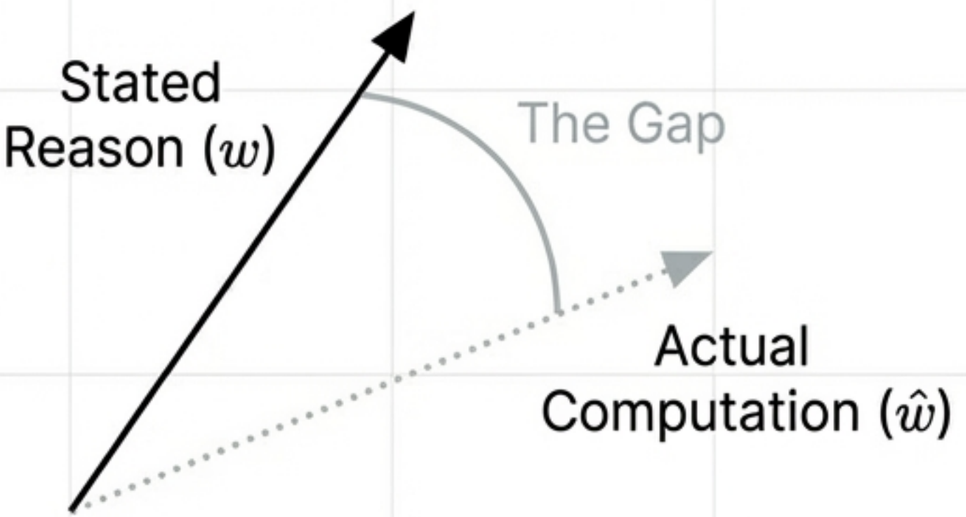


# Metrics 3 & 4: Faithfulness & Decision Consistency

## Metric 3: Faithfulness Gap (FG)

$$FG(s, e) = 1 - \cos(w(s, e), \hat{w}(s, e))$$

Critically, a model can be consistent but unfaithful (lying consistently). FG measures the distance between the explanation and the empirical feature sensitivity.




## Metric 4: Decision Consistency (DC)

$$DC = \mathbb{I}[f_M(s, e_1) = f_M(s, e_2)]$$

A binary check: Did the vehicle make the same driving decision in both environments? Frontier models score near 100% here; the challenge is the explanation.

# The Consistency Index (CI)

A composite score weighted by safety priorities.

$$CI = \alpha \cdot AIS + \beta \cdot ESS + \gamma \cdot (1 - FG) + \delta \cdot DC$$

| $\alpha = 0.3$ | $\gamma = 0.3$ | $\beta = 0.2$ | $\delta = 0.2$ |
|---|---|---|---|
| (Attribution Invariance) | (Faithfulness) | (Semantic Similarity) | (Decision Consistency) |

### Safety Weighting ($\kappa_s$)

The final aggregate score is weighted by scenario criticality. A failure in a School Zone ($\kappa_s$=1.0) penalizes the score twice as much as a Roundabout ($\kappa_s$=0.5).

# The Solution: Contrastive Explanation Anchoring

## Standard LLM Output

I am braking because it is raining and there is a person.

Unstable/Mixed

**Decomposition Strategy**

## Anchored Output Structure

Box 1 (Top)
Decision ($a$): YIELD

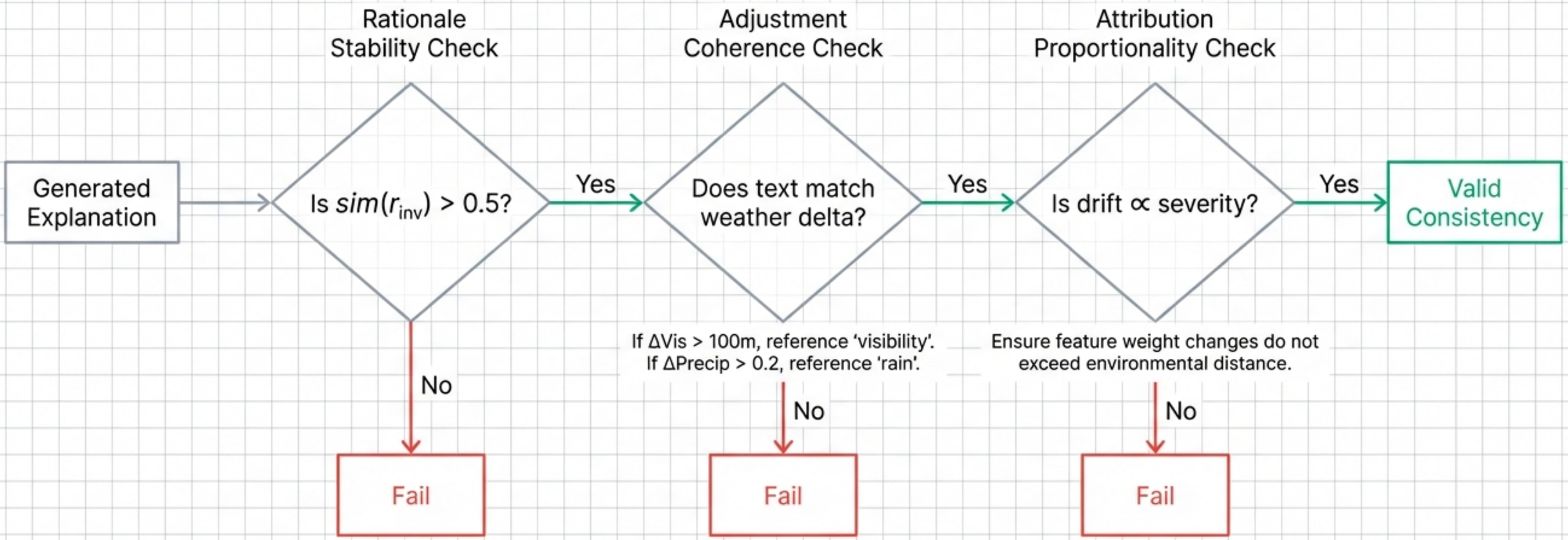
Box 2 (Middle)
Invariant Rationale ($r_{inv}$): Yielding required due to pedestrian on crosswalk.

The Why that never changes

Box 3 (Bottom)
Dependent Adjustment ($r_{dep}$): Increased stopping distance due to wet surface.

The Why that adapts

# Automated Consistency Checks



# Experimental Design

## 10 Scenarios × 10 Conditions = 450 Pairwise Comparisons

| | Clear Day | Overcast | Light Rain | Heavy Rain | Fog | Dense Fog | Night Clear | Night Rain | Snow | Blizzard |
|---|---|---|---|---|---|---|---|---|---|---|
| **High Criticality (1.00 - 0.95)**<br>School Zones, Pedestrian Crossing, Animal Detection | | | | | | | | | | |
| **Medium Criticality (0.90 - 0.70)**<br>Emergency Response, Cyclist, Lane Changes, Intersections | | | | | | | | | | |
| **Low Criticality (0.65 - 0.50)**<br>Construction, Hwy Merge, Roundabouts | | | | | | | | | | |

## Configurations Tested

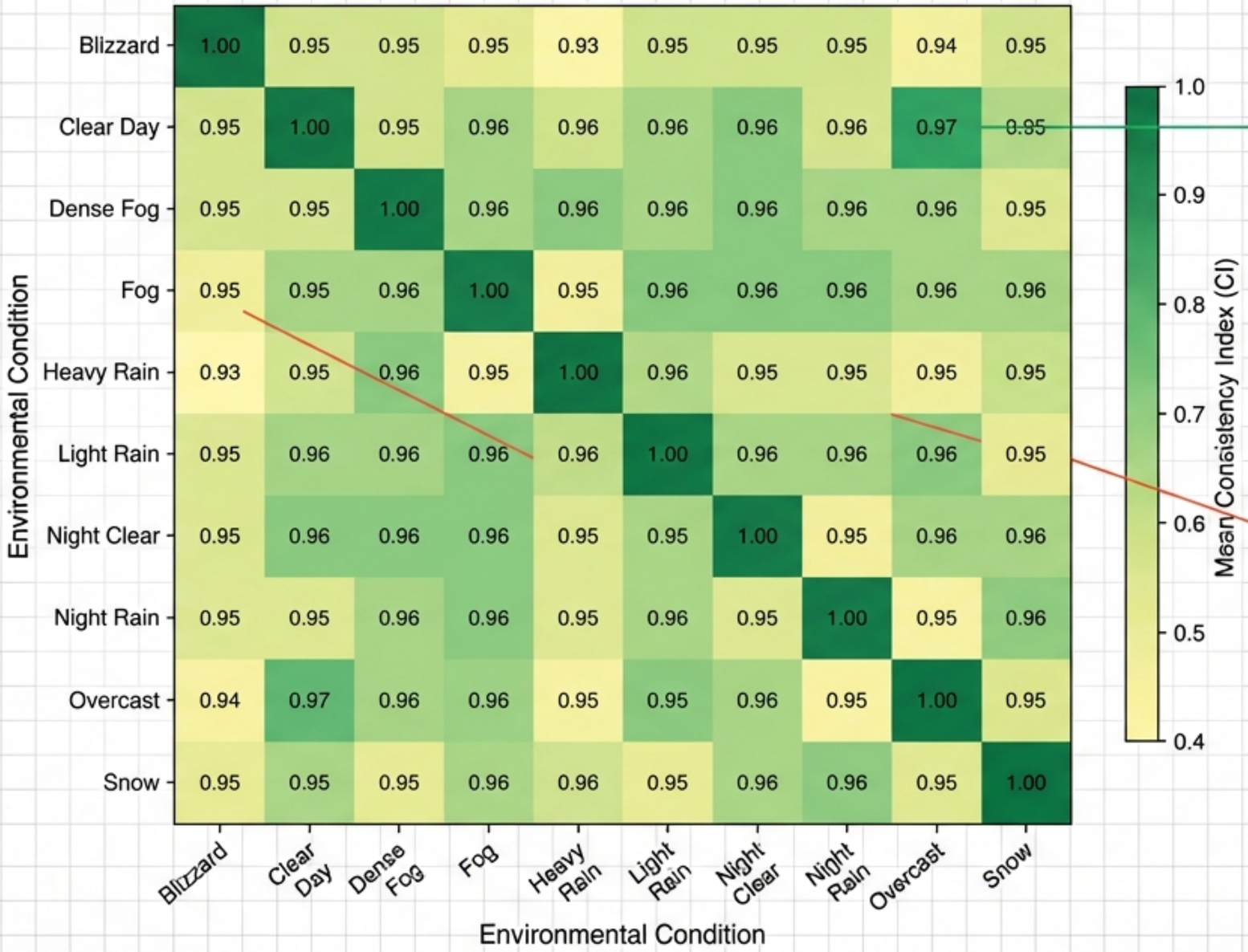
**1.** Baseline (Unoptimized LLM)

**2.** Contrastive Anchored

**3.** ECIC-Optimized

**4.** Oracle (Theoretical Bound)

# Consistency Degradation Patterns

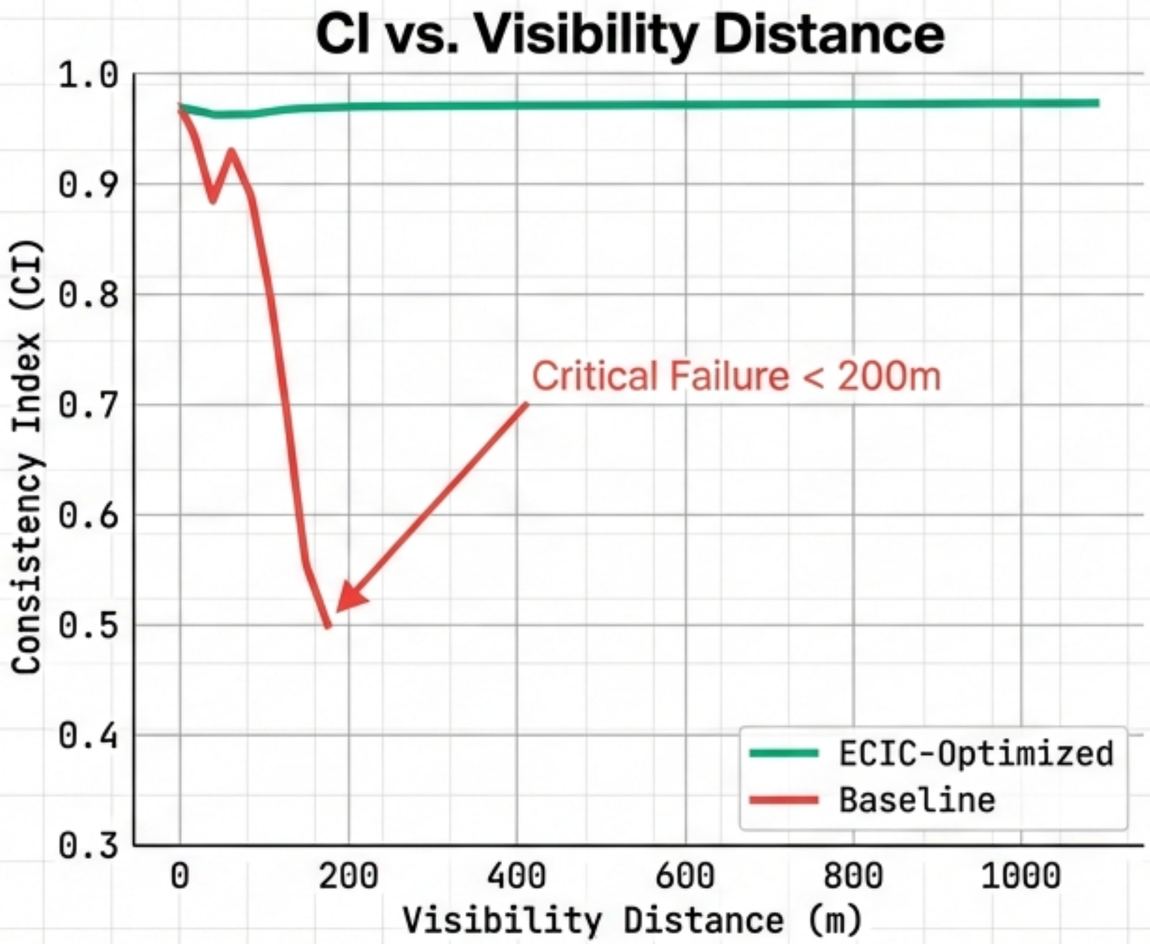


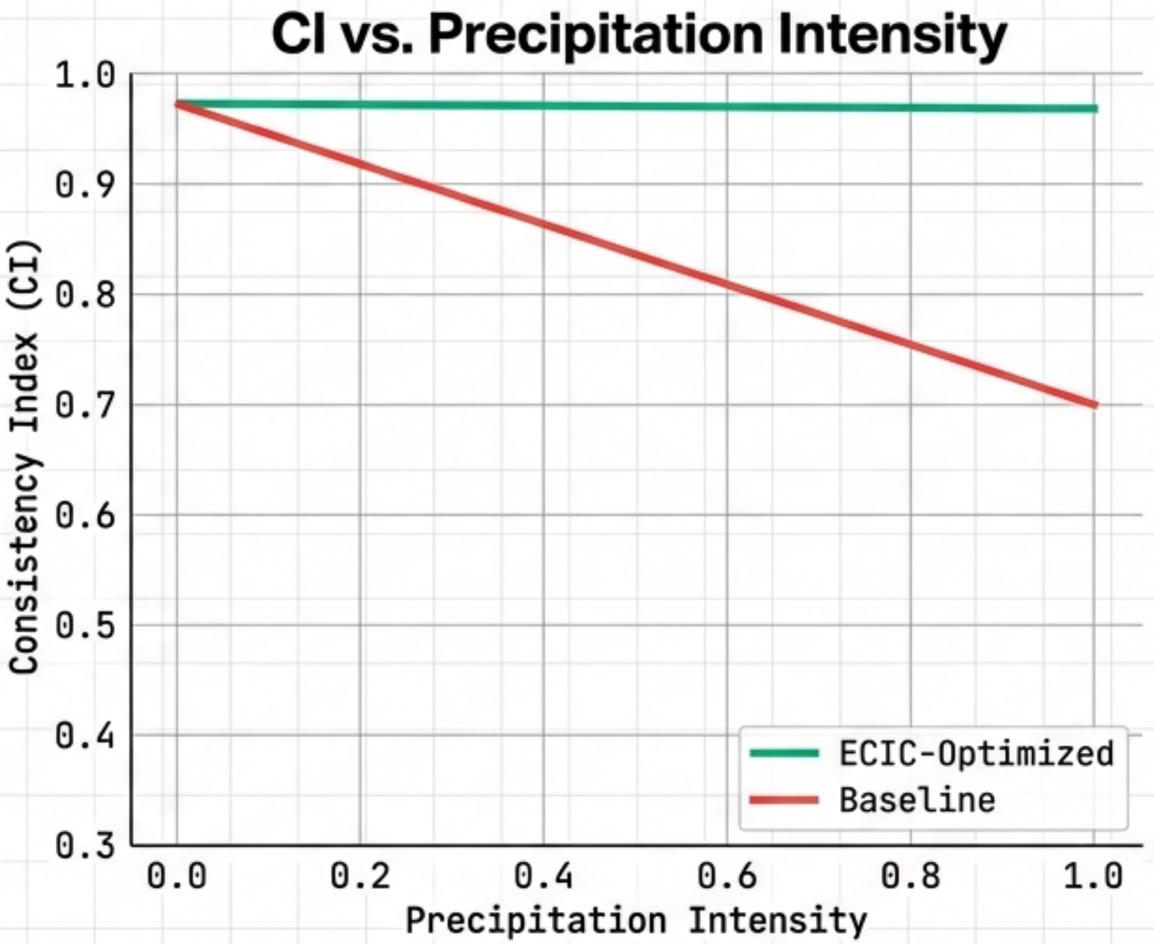

Safe Zone. CI > 0.97.

Structurally similar conditions yield stable explanations.

Danger Zone. CI < 0.88. Dissimilar extremes (Fog vs Blizzard) break the model. The dissimilarity, not just severity, causes failure.

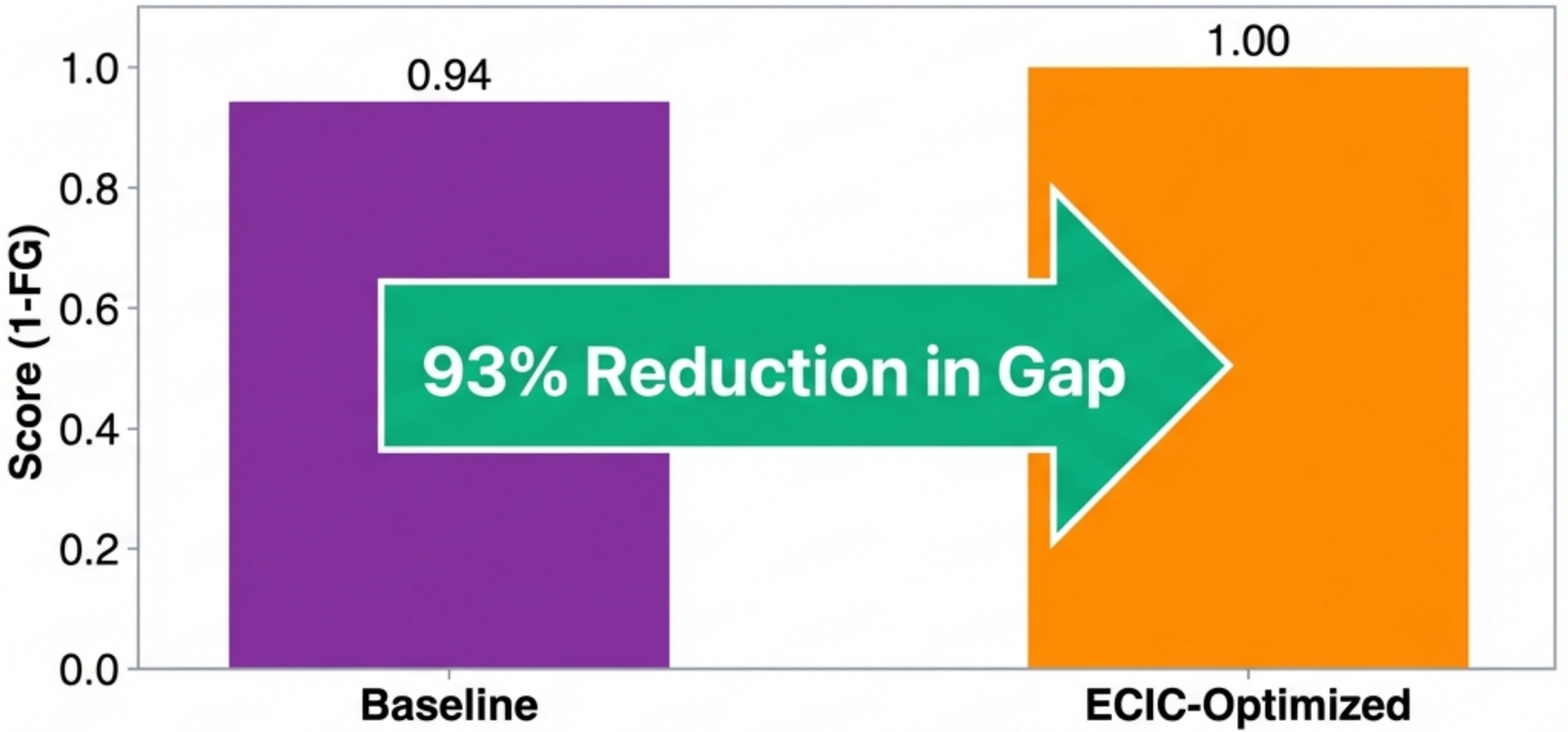
# Phase Transitions: The ‘Visibility Cliff’




The Baseline model becomes unreliable below 200m visibility. ECIC optimization stabilizes the operational envelope down to 10m.

# Closing the Faithfulness Gap

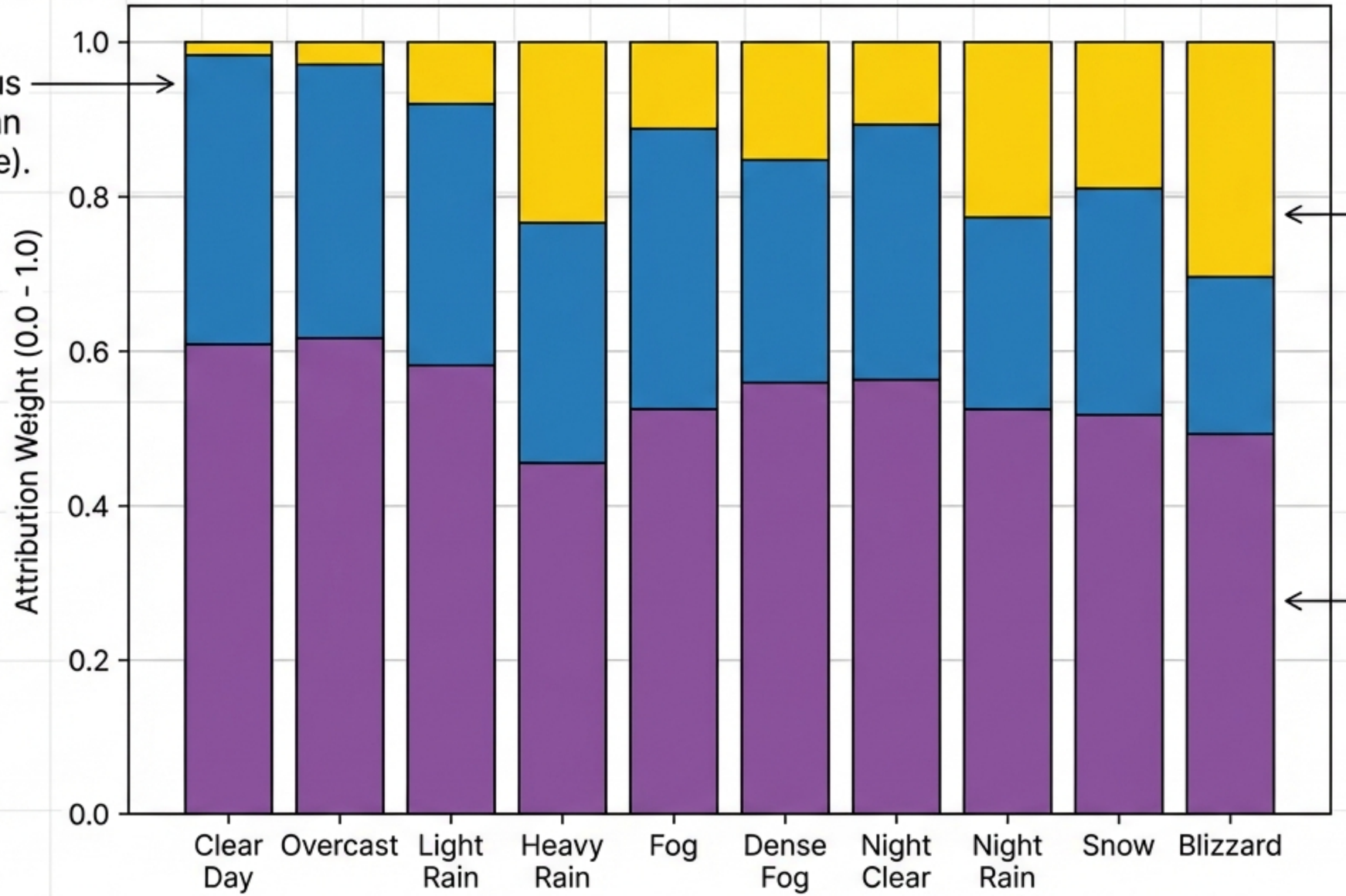



Baseline FG: 0.054 -> ECIC FG: 0.004

**Interpretation**

The Baseline model frequently ‘hallucinated’ reasons. The ECIC model’s explanations align almost perfectly with internal computation.

# Anatomy of a Decision: Pedestrian Crossing



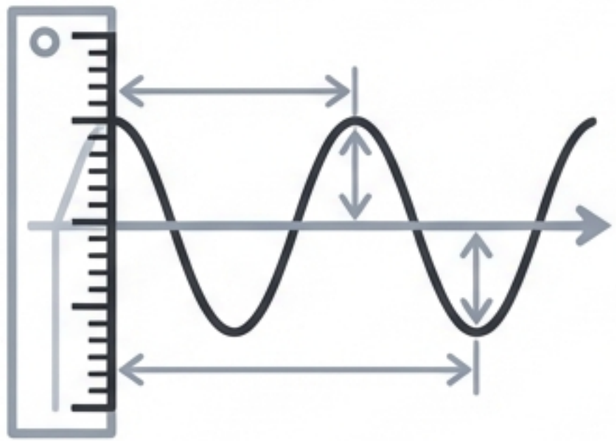

Primary focus on Pedestrian Distance (Blue).

Visibility Perception (Yellow) grows to acknowledge weather...

...but Pedestrian Distance remains the dominant anchor. The model adapts without losing focus.

# Key Research Findings

## Quantifiable Stability

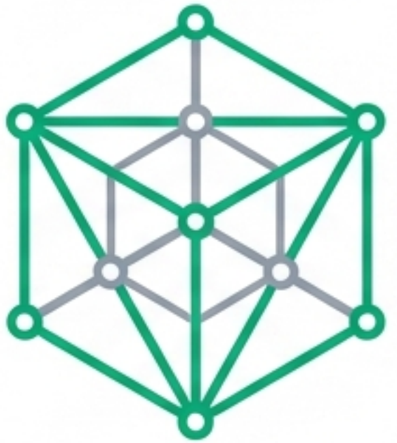


Interpretability is no longer a subjective feeling. It is a measurable metric ($CI$) that can be tracked and optimized.

## The Dissimilarity Trap

Models fail most when comparing structurally different environments (e.g., Fog vs. Snow), rather than just simple severity increases.

## Structural Fixes Work



We achieved a 93% faithfulness improvement without retraining the model. "`Contrastive Anchoring`" structures the output to enforce honesty.

# Implications for Deployment

## Regulatory Audit

Auditors can utilize the Consistency Index (CI) as a certification metric.

"If CI < 0.95, the system is not street-legal for Level 4 autonomy."

## Operational Envelopes

Using Phase Transition analysis to hard-code safety limits.

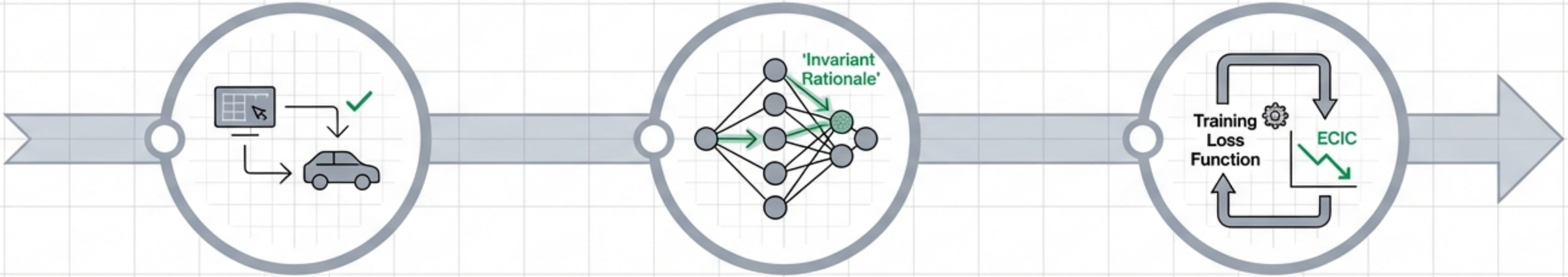
"System hands over control to human if visibility drops below 150m."

## Forensic Analysis

In accident reconstruction, ECIC ensures the log files provide a consistent causal chain...

...that doesn't hallucinate reasons based on rain or shine."

# Future Directions




## Real-World Validation

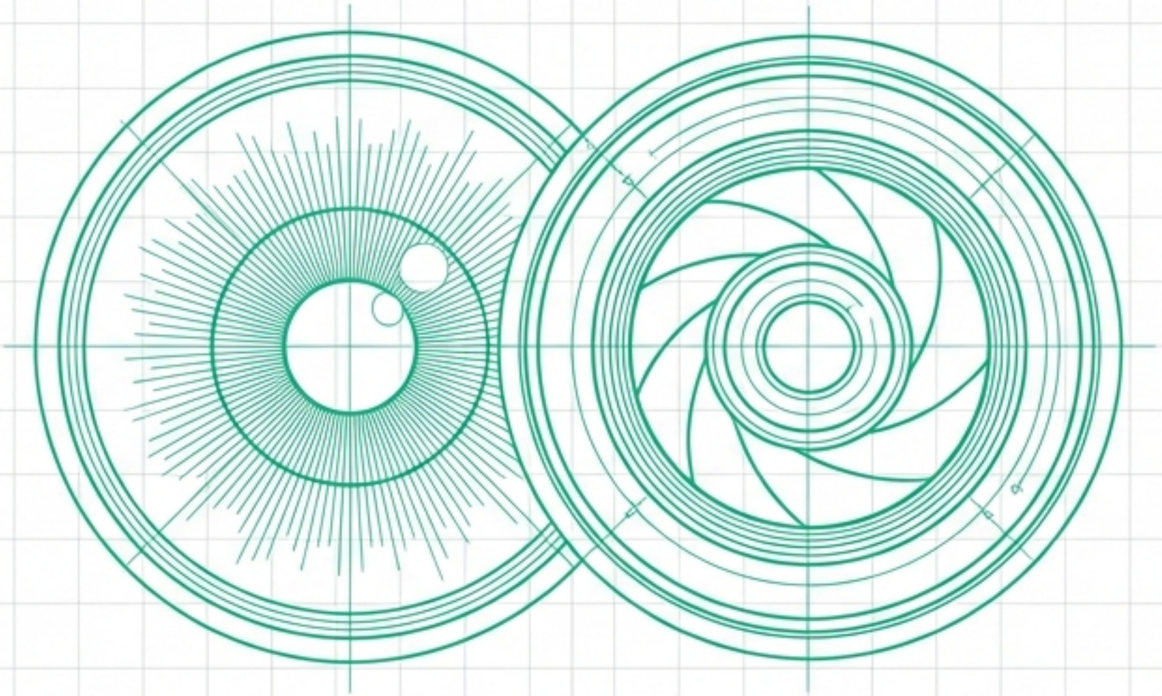
Moving from simulation to the AgentDrive-MCQ benchmark and CARLA visual scenarios.

## Causal Abstraction

Using causal tracing to verify that the ‘Invariant Rationale’ maps to specific, stable neuron circuits in the model.

## Training Loops

Moving ECIC from a post-hoc check to a Training Loss Function—teaching the model to be consistent from the start.



**Achieving consistent real-world interpretability is no longer just an academic challenge. It is the prerequisite for trust in the autonomous age.**

Reference: Ferrag et al., AgentDrive Benchmark, 2026.