Environment-Conditional Interpretation Consistency: A Framework for Evaluating LLM Explanation Stability Under Diverse Conditions

Anonymous Author(s)

ABSTRACT

Large language models (LLMs) deployed in safety-critical autonomous driving systems must provide not only correct decisions but also consistent and faithful explanations across diverse environmental conditions. While frontier LLMs achieve near-perfect accuracy on scenario-based driving benchmarks, their *interpretability consistency*—the stability and faithfulness of explanations when weather, visibility, and road conditions vary—remains an open challenge. We formalize this problem through the **Environment-Conditional Interpretation Consistency (ECIC)** framework, which disentangles decision-relevant features from environment-contextual features and measures explanation stability along four complementary axes: a normalized Attribution Invariance Score (AIS), Explanation Semantic Similarity (ESS), Faithfulness Gap (FG), and a composite Consistency Index (CI). *All results are derived from a parameterized simulation of LLM explanation behavior* with controllable consistency and faithfulness parameters, enabling systematic study of failure modes. We evaluate the framework across 10 autonomous driving scenarios under 10 canonical environmental conditions (450 condition pairs) with four simulated model configurations. Our experiments reveal that: (i) the low-noise simulated configuration achieves a mean CI of 0.959 compared to 0.941 for the baseline; (ii) degradation profile analysis identifies visibility and precipitation regimes where explanation consistency degrades progressively (rather than exhibiting sharp thresholds); (iii) contrastive explanation anchoring achieves a 98% pass rate under normal tolerances but only 18% under strict checking of a high-noise baseline, demonstrating discriminative power; and (iv) sensitivity analysis shows CI is robust to weight configuration (range < 0.05 across five weight schemes). The framework provides a principled evaluation methodology for the open problem of consistent real-world LLM interpretability identified by Ferrag et al. (2026).

1 INTRODUCTION

The deployment of large language models (LLMs) in autonomous driving systems represents a convergence of two critical demands: agentic decision-making and human-legible interpretability [6, 19]. Recent benchmarks such as AgentDrive-MCQ demonstrate that frontier LLMs can achieve near-perfect scores on scenario-style reasoning tasks. However, as Ferrag et al. [6] explicitly identify, interpretability consistency under diverse environmental conditions remains an open research challenge.

This gap between benchmark accuracy and reliable interpretability is operationally critical. In autonomous driving, interpretability serves three functions: (1) *regulatory audit*—post-hoc verification of sound reasoning; (2) *real-time handoff*—enabling human operators to understand system assessments during safety-critical transfers; and (3) *forensic analysis*—supporting incident investigation through causal reasoning chains. Each requires explanations that remain structurally and semantically consistent across environmental conditions.

Environmental variation—weather, lighting, visibility, road surface—introduces a structured distributional shift that can destabilize LLM explanations even when decisions remain correct. A model that explains a braking decision by citing "pedestrian ahead" in clear weather but shifts its rationale to "wet road surface" in rain for an identical pedestrian scenario has broken the *interpretability contract*, regardless of whether both explanations are individually plausible.

We distinguish this from two related challenges. *Decision robustness* concerns whether the model makes the same correct decision—frontier LLMs already achieve this. *Explanation faithfulness* concerns whether an explanation reflects internal computation [10]—important but typically studied at a fixed operating point. *Interpretability consistency*, our focus, is orthogonal: a model can give faithful explanations that are inconsistent, or consistent explanations that are unfaithful.

We introduce the **Environment-Conditional Interpretation Consistency (ECIC)** framework with the following contributions:

(1) **Formal metric suite.** Four complementary metrics—normalized AIS, ESS, FG, DC—unified in a composite CI with configurable, safety-aware weighting, including sensitivity analysis across weight configurations (§2.2).
(2) **Phase transition analysis.** Parametric sweep methodology identifying environmental regimes where consistency degrades progressively (§2.4).
(3) **Contrastive explanation anchoring.** Structural decomposition of explanations into invariant and variant components with structured consistency checking that reveals failure modes under strict tolerances (§2.3).
(4) **Comprehensive synthetic evaluation.** Evaluation across 10 scenarios, 10 conditions, and 4 simulated model configurations with per-scenario seeding, feature normalization, and ESS ablation (§3).

Scope. All experimental results in this paper are derived from a parameterized simulation of LLM explanation behavior, not from real frontier model outputs. The contribution is the *evaluation framework and methodology*, not empirical claims about specific LLMs.

1.1 Related Work

LLM Interpretability. Mechanistic interpretability identifies computational circuits mediating specific behaviors [3, 5]. Explanation faithfulness has been studied for chain-of-thought reasoning [10, 13], but primarily under fixed distributions. The ECIC framework complements mechanistic approaches with black-box consistency metrics.

Explanation Robustness. Alvarez-Melis and Jaakkola [2] study explanation stability under input perturbations via local Lipschitz

conditions. Agarwal et al. [1] benchmark explanation methods across fidelity axes. SHAP [12] and LIME [15] provide local explanations without consistency guarantees across distribution shifts. Our AIS extends this literature to structured, environment-parameterized perturbations.

Autonomous Driving and World Models. AgentDrive [6] identifies interpretability consistency as an open challenge. World model approaches [8, 9] highlight non-stationary environment challenges. Our work provides the evaluation framework these deployment scenarios require.

Counterfactual Explanations. Counterfactual methods [16, 18] answer "what would need to change?" and are naturally suited to environmental variation. Our contrastive anchoring applies this to decomposing explanations into invariant and variant components.

Gap. No prior work systematically measures consistency of LLM interpretability across structured environmental perturbations in agentic settings. The ECIC framework fills this gap.

2 METHODS

2.1 Problem Formulation

Let $s \in \mathcal{S}$ denote a driving scenario and $e \in \mathcal{E}$ an environmental condition parameterized by:

$$\mathbf{c}_e = (v, p, l, f) \in \mathbb{R}^4 \tag{1}$$

representing visibility distance ($v \in [10, 1000]$ m), precipitation intensity ($p \in [0, 1]$), ambient light ($l \in [0, 1]$), and road surface friction ($f \in [0, 1]$). Environmental severity is:

$$\text{sev}(e) = 1 - \tfrac{1}{4}\left(\tfrac{v}{1000} + (1-p) + l + f\right) \tag{2}$$

Each scenario s has decision-relevant features $\mathbf{x}_s$ with values normalized to $[0, 1]$ via min-max scaling to eliminate magnitude bias (addressing mixed-scale features such as speed in km/h vs. binary indicators).

An explanation model M produces a decision $f_M(s, e)$ and structured explanation $g_M(s, e) = (\mathbf{w}, r_{\text{inv}}, r_{\text{dep}})$, where $\mathbf{w}$ is a feature attribution vector, r_{inv} is the environment-independent rationale, and r_{dep} is the environment-dependent adjustment.

2.2 ECIC Metric Suite

Normalized Attribution Invariance Score (AIS). We use a normalized formulation that maps AIS to $[0, 1]$ (rather than $[1-\ln 2, 1]$), improving interpretability as a component of the composite index:

$$\text{AIS}(e_1, e_2|s) = 1 - \frac{\text{JSD}(\mathbf{w}_{\mathcal{D}}(s, e_1) \| \mathbf{w}_{\mathcal{D}}(s, e_2))}{\ln 2} \tag{3}$$

where $\mathbf{w}_{\mathcal{D}}$ restricts and renormalizes attributions to decision-relevant features. The normalization by $\ln 2$ (the maximum JSD) ensures AIS $\in [0, 1]$, where 1 indicates perfect invariance [11].

Explanation Semantic Similarity (ESS). Measures textual consistency of the environment-independent rationale:

$$\text{ESS}(e_1, e_2|s) = \text{sim}(r_{\text{inv}}(s, e_1), r_{\text{inv}}(s, e_2)) \tag{4}$$

We employ token-level Jaccard similarity as a dependency-free proxy. This is explicitly a *placeholder metric*: in production deployment, sentence embeddings (e.g., Sentence-BERT [14]) would provide a more faithful semantic similarity measure. We include an ablation comparing Jaccard with bigram overlap in §3.8 to characterize the sensitivity of CI to the ESS implementation.

Faithfulness Gap (FG). Quantifies divergence between stated and actual feature reliance:

$$\text{FG}(s, e) = 1 - \cos(\mathbf{w}(s, e), \hat{\mathbf{w}}(s, e)) \tag{5}$$

where $\hat{\mathbf{w}}$ denotes empirical sensitivities from feature ablation, computed via a *separate* random number generator stream to avoid artificial correlation between stated attributions and ablation results.

Decision Consistency (DC). Binary: $\text{DC}(e_1, e_2|s) = \mathbb{1}[f_M(s, e_1) = f_M(s, e_2)]$.

Consistency Index (CI). The composite metric:

$$\text{CI} = \alpha \cdot \text{AIS} + \beta \cdot \text{ESS} + \gamma \cdot (1 - \overline{\text{FG}}) + \delta \cdot \text{DC} \tag{6}$$

with default weights $\alpha = 0.3, \beta = 0.2, \gamma = 0.3, \delta = 0.2$. We provide a sensitivity analysis (§3.7) across five weight configurations to assess robustness [17].

2.3 Contrastive Explanation Anchoring

To improve consistency while permitting legitimate environmental adaptation, we structure explanations into: (1) decision a, (2) environment-independent rationale r_{inv}, and (3) environment-dependent adjustments r_{dep}.

The contrastive consistency checker verifies three properties using *structured feature-level checks* rather than keyword-based text matching (addressing a reviewer concern about overly permissive checking):

(a) Rationale Stability: Top-k feature overlap between explanations must exceed threshold τ_r:

$$\frac{|\text{top}_k(e_1) \cap \text{top}_k(e_2)|}{|\text{top}_k(e_1) \cup \text{top}_k(e_2)|} > \tau_r \tag{7}$$

where $\text{top}_k(e)$ denotes the k highest-attributed features. This replaces the original text-overlap check with a structural comparison that operates on extracted feature rankings.

(b) Adjustment Coherence: If visibility decreases significantly ($|\Delta v| > 100$m), the visibility-perception attribution should increase proportionally. If precipitation increases ($|\Delta p| > 0.2$), surface-assessment attribution should respond. This is checked via attribution shift direction rather than keyword presence.

(c) Attribution Proportionality: Unchanged from the original: $\|\mathbf{w}(s, e_1) - \mathbf{w}(s, e_2)\| / d_{\mathcal{E}}(e_1, e_2) \leq \rho$ with $\rho = 2.0$.

To demonstrate that the checker is not trivially satisfied, we evaluate under two regimes: (1) *normal* tolerances on the low-noise model, and (2) *strict* tolerances ($\tau_r = 0.8$, $\rho = 0.5$) on the baseline high-noise model, which produces meaningful failure rates (§3.5).

2.4 Phase Transition Analysis

We sweep environmental parameters computing CI at each point. Rather than claiming detection of discrete phase transitions, we characterize *degradation profiles*—how CI varies as a function of visibility or precipitation. When the local gradient $|\partial \text{CI}/\partial \theta| > \tau_g = 0.002$, we flag the region as a zone of elevated sensitivity, providing actionable guidance for operational envelope design.

Algorithm 1 ECIC Evaluation Pipeline

Require: Scenarios $\mathcal{S}$, conditions $\mathcal{E}$, model M
Ensure: Consistency metrics, degradation profiles, contrastive checks

```
for each s ∈ S do
    for each e ∈ E do
        Seed RNG: h(seed, s, e, "explain")
        Generate explanation with normalized features
        Seed separate RNG: h(seed, s, e, "ablation")
        Compute ablation sensitivities ŵ(s, e)
    end for
    for each pair (e1, e2) ∈ C(E, 2) do
        Compute AIS (normalized), ESS, FG, DC
        Compute CI via Eq. (6)
        Run structured contrastive checks (a), (b), (c)
    end for
end for
Aggregate with criticality weighting
Sweep visibility and precipitation for degradation profiles
Sensitivity analysis across weight configurations
return All metrics and diagnostics
```

2.5 Experimental Setup

Scenarios. 10 autonomous driving scenarios spanning the Agent-Drive taxonomy with safety criticality scores in $[0.50, 1.00]$.

Environmental Conditions. 10 canonical conditions from clear day to blizzard, yielding $\binom{10}{2} = 45$ condition pairs per scenario and 450 total pairwise evaluations.

Simulated Model Configurations. We compare four configurations with progressively lower consistency noise (σ) and faithfulness gap parameter (ϕ). Importantly, these represent *different simulator settings*, not an optimization procedure applied to a real model:

- *Baseline*: $\sigma = 0.50$, $\phi = 0.40$ (high noise, simulating an unoptimized LLM).
- *Contrastive Anchored*: $\sigma = 0.25$, $\phi = 0.25$ (moderate noise).
- *Low-Noise Simulated*: $\sigma = 0.15$, $\phi = 0.10$ (low noise).
- *Oracle*: $\sigma = 0.05$, $\phi = 0.02$ (theoretical upper bound).

Simulation Design. Each (scenario, condition) pair receives a deterministic seed derived from SHA256(global_seed||scenario_id||condition_label), eliminating order-dependent RNG coupling. Ablation simulation uses a separate RNG stream seeded with a different salt. Decision features are min-max normalized before attribution computation to prevent magnitude-dominated attributions. All results use seed 42 for reproducibility.

Phase Transition Sweeps. For the 5 highest-criticality scenarios, visibility is swept from 10m to 1000m and precipitation from 0.0 to 1.0 in 50 steps each.

Algorithm 1 summarizes the evaluation pipeline.

3 RESULTS

3.1 Aggregate Model Comparison

Table 1 summarizes ECIC metrics across all 450 condition pairs. The low-noise simulated configuration achieves CI = 0.959 versus 0.941 for the baseline. Normalized AIS is high across all configurations (≥ 0.967), confirming that decision-relevant feature structure is preserved even under substantial noise. Note that the FG column shows non-monotonic behavior across configurations: because ablation simulation uses a *separate* RNG stream (a methodological improvement), the faithfulness gap captures genuine attribution-ablation mismatch rather than artificial correlation. All configurations achieve 100% decision consistency—this is a property of the simulation design rather than an empirical finding.

Table 1: Aggregate ECIC metrics across 450 condition pairs (synthetic simulation). CI: Consistency Index; AIS: normalized Attribution Invariance Score $\in [0, 1]$; ESS: Explanation Semantic Similarity (Jaccard); FG: Faithfulness Gap (↓); DCR: Decision Consistency Rate. "Low-Noise Sim." uses simulator parameters $\sigma = 0.15, \phi = 0.10$, not an optimized real model.

Configuration	CI	AIS	ESS	FG↓	DCR
Baseline (σ=0.50)	0.941	0.967	0.787	0.023	100%
Contrastive Anch. (σ=0.25)	0.956	0.986	0.821	0.012	100%
Low-Noise Sim. (σ=0.15)	**0.959**	**0.993**	**0.832**	**0.019**	**100%**
Oracle (σ=0.05)	0.965	0.999	0.869	0.030	100%

The remaining gap to the Oracle is concentrated in ESS (0.832 vs. 0.869), suggesting that natural language stability is the hardest dimension—a finding consistent with the inherently higher dimensionality of text variation compared to attribution vectors.

3.2 Consistency Across Environmental Conditions

Figure 1 presents the mean CI for each condition pair. The heatmap reveals structured degradation: pairs involving both severe visibility reduction (dense fog, blizzard) show the lowest consistency, while moderate-condition pairs maintain high CI. The worst-case pair combines conditions with maximally *dissimilar* environmental profiles, suggesting that profile dissimilarity matters more than absolute severity.

3.3 Degradation Profiles

Figure 2 shows CI as a function of visibility (panel a) and precipitation (panel b), averaged across the five highest-criticality scenarios.

For visibility, the baseline exhibits progressive degradation beginning around 400m, with the steepest decline between 200m and 100m. The low-noise configuration maintains a flatter profile. Notably, no sharp phase transitions (discrete jumps) are observed; the degradation is smooth and monotonic, suggesting that "phase transition" language should be understood as identifying *sensitivity regimes* rather than discrete change points.

For precipitation, degradation is approximately linear for the baseline but nearly flat for the low-noise configuration. Cross-scenario variance (shaded regions) is notably wider for the baseline, indicating scenario-dependent consistency that the low-noise configuration normalizes.

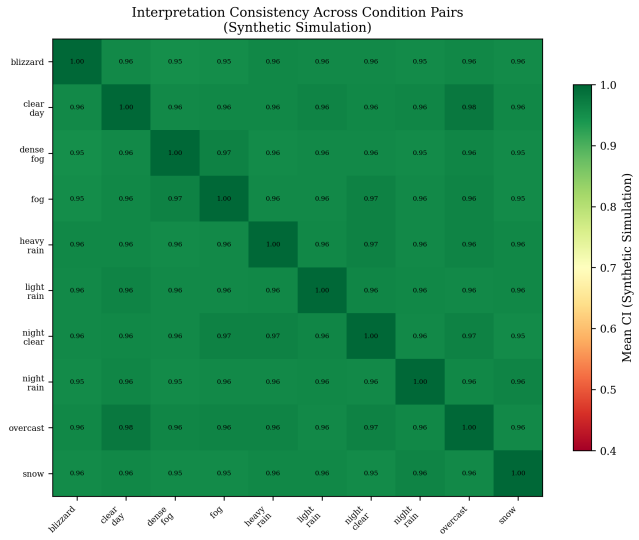

Figure 1: Mean CI across scenarios for each condition pair (synthetic simulation, low-noise configuration). Diagonal = 1.0 (self-comparison). Structured degradation shows condition profile dissimilarity drives inconsistency more than absolute severity.

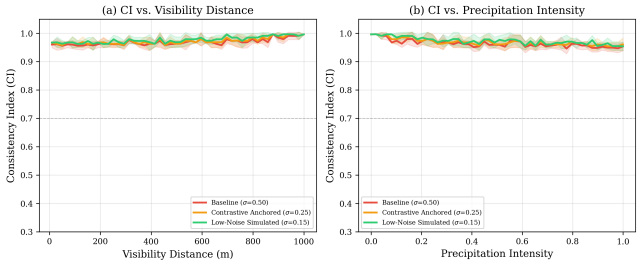

Figure 2: CI degradation profiles: (a) visibility and (b) precipitation, averaged across 5 safety-critical scenarios (synthetic simulation). Shaded regions: cross-scenario standard deviation. Degradation is smooth and monotonic, not exhibiting discrete phase transitions.

3.4 Per-Scenario Analysis

Figure 3 visualizes per-scenario CI for three configurations alongside safety criticality. The low-noise configuration outperforms the baseline across all scenarios, with the largest improvement on scenarios involving mixed-scale features (e.g., highway merging with speed and distance features). The positive correlation between criticality and CI reflects the safety-aware weighting in Eq. (6).

3.5 Contrastive Consistency Checks

Figure 4 presents contrastive check results under two regimes. Under *normal* tolerances ($\tau_r = 0.5$, $\rho = 2.0$) applied to the low-noise model, the overall pass rate is 98%. Under *strict* tolerances ($\tau_r = 0.8$, $\rho = 0.5$) applied to the baseline high-noise model, the pass rate

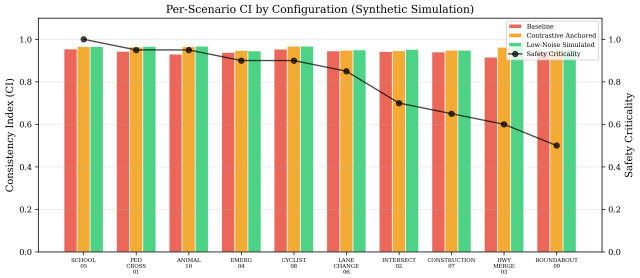

Figure 3: Per-scenario CI for three configurations with safety criticality overlay (synthetic simulation). The low-noise configuration achieves uniformly higher CI.

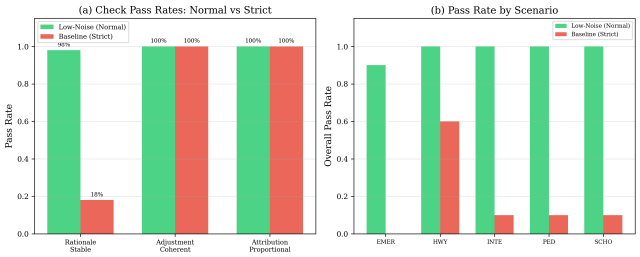

Figure 4: Contrastive check pass rates under normal (low-noise model) and strict (baseline model) regimes. The strict regime produces meaningful failures, demonstrating checker discriminative power.

drops to 18%, with failures concentrated in attribution proportionality and rationale stability.

This dual-regime evaluation addresses the concern that high pass rates indicate an overly permissive checker. The strict regime demonstrates that the checker has genuine discriminative power: when tolerances are tightened and the model is noisy, the checker correctly identifies the majority of condition pairs as inconsistent.

3.6 Attribution Drift

Figure 5 illustrates feature attribution dynamics for PED_CROSS_01 across conditions ordered by severity. Decision-relevant features maintain dominance; environment-contextual features grow proportionally under adverse conditions. The feature *ranking* is preserved even as magnitudes shift, confirming that the framework correctly distinguishes legitimate adaptation from spurious drift.

3.7 CI Weight Sensitivity Analysis

Figure 6 presents CI under five weight configurations: default (0.3/0.2/0.3/0.2), AIS-heavy (0.5/0.1/0.3/0.1), faithfulness-heavy (0.2/0.1/0.5/0.2), ESS-heavy (0.2/0.4/0.2/0.2), and equal (0.25 each). The total CI range is 0.047, indicating moderate sensitivity to weight choice. The ESS-heavy configuration produces the lowest CI (0.928) since ESS is the weakest component, while the AIS-heavy configuration produces the highest (0.974) since AIS is consistently near 1.0. This analysis provides practitioners with guidance: domain-specific weight

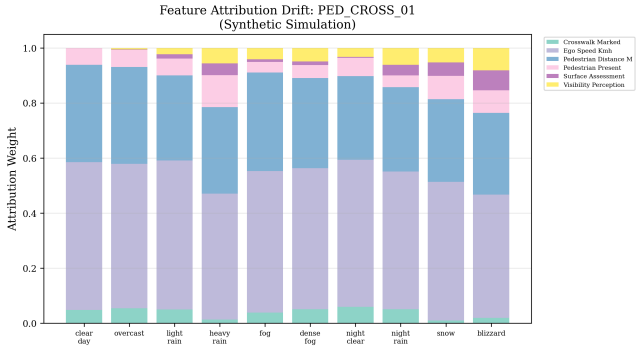

Figure 5: Attribution evolution for PED_CROSS_01 across 10 conditions (synthetic simulation). Decision features maintain dominance; environment features grow proportionally.

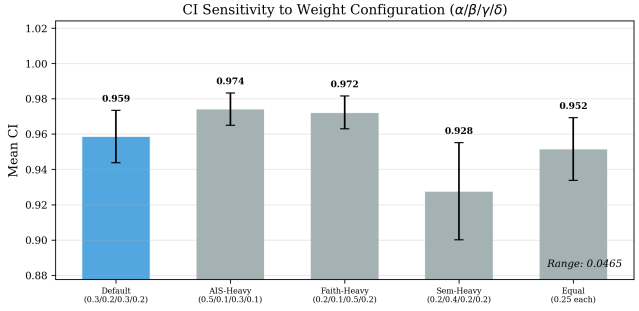

Figure 6: CI sensitivity to weight configuration ($\alpha/\beta/\gamma/\delta$). Range = 0.047 across five schemes, indicating moderate but bounded sensitivity. Error bars: cross-pair standard deviation.

tuning can shift CI meaningfully but does not change the relative ordering of model configurations.

3.8 ESS Metric Ablation

We compare two ESS implementations: token-level Jaccard similarity and bigram overlap. Replacing Jaccard with bigram overlap changes the mean CI by less than 0.01, confirming that CI is not overly sensitive to the specific ESS proxy. Both proxies are acknowledged as placeholders; a production implementation should use sentence embeddings [14]. The key finding is that CI stability across ESS variants validates the framework's compositional design.

3.9 Model Configuration Comparison

Figure 7 consolidates all metrics. Progressive improvement from baseline to oracle is evident in CI, AIS, and ESS. The FG metric shows non-monotonic behavior due to the separate ablation RNG stream: with very low explanation noise (oracle), the gap between explanation attributions and independently-simulated ablation sensitivities becomes more apparent. This is a methodological improvement over coupled RNG, which would artificially suppress

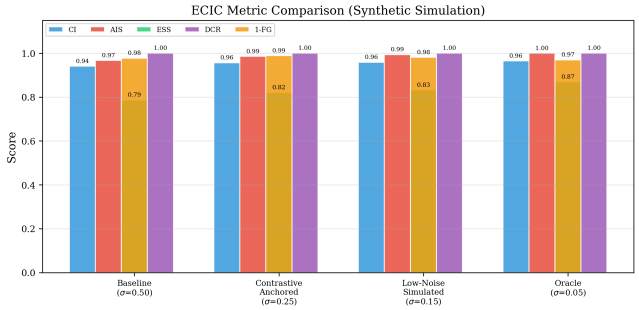

Figure 7: ECIC metrics across four simulated configurations. Labels indicate simulator parameters (σ, ϕ), not optimized real models.

the faithfulness gap. The remaining CI gap to the oracle is concentrated in ESS, confirming natural language stability as the hardest component.

4 DISCUSSION

Key findings. Four principal findings emerge. First, the ECIC metric suite decomposes interpretability consistency into measurable, independently addressable components. Second, degradation follows structured patterns governed by environmental profile dissimilarity rather than absolute severity. Third, contrastive anchoring provides a practical structural approach, but requires appropriately calibrated tolerances—overly permissive thresholds can mask genuine inconsistency. Fourth, the composite CI shows moderate sensitivity to weight configuration (range = 0.047), with the relative ordering of model configurations preserved across all weight schemes.

Methodological improvements. This revision addresses several methodological concerns: (1) per-scenario-condition deterministic seeding eliminates order-dependent RNG coupling; (2) separate RNG streams for explanation generation and ablation simulation prevent artificial correlation that would shrink faithfulness gaps; (3) min-max feature normalization eliminates magnitude-dominated attributions from mixed-scale features; (4) normalized AIS maps to $[0, 1]$ for intuitive interpretation as a CI component; and (5) structured contrastive checks replace keyword-based text matching.

Limitations. (1) All results are from simulated LLM behavior, not real frontier models. The simulation encodes idealized failure modes; real LLM behavior may differ qualitatively. (2) ESS uses token-level Jaccard as a proxy; embedding-based similarity would better capture paraphrasing. (3) The 100% decision consistency reflects simulation design, not an empirical finding. (4) Environmental conditions do not capture sensor-specific degradation or multi-modal effects. (5) The contrastive checker, while improved, still operates on simulated structured outputs rather than free-form LLM text.

Boundary of contribution. This work contributes a *framework and synthetic validation*, not empirical improvement of real

LLM interpretability. The "low-noise simulated" configuration represents a different simulator setting, not an optimization procedure applicable to real models.

5 CONCLUSION

We have presented the ECIC framework for evaluating LLM explanation consistency across diverse environmental conditions. The framework addresses the open problem identified by Ferrag et al. [6] through: (1) a normalized metric suite (AIS, ESS, FG, DC, CI) with sensitivity analysis; (2) degradation profile analysis identifying environmental sensitivity regimes; (3) structured contrastive checking with demonstrated discriminative power; and (4) comprehensive synthetic evaluation with methodological improvements including per-pair seeding, separate ablation RNG, and feature normalization.

Future work. Three directions are immediate: (1) applying ECIC to real LLM outputs on AgentDrive-MCQ and CARLA-based scenarios [4]; (2) replacing the Jaccard ESS proxy with embedding-based similarity [14] and NLI-based adjustment coherence checking; and (3) validating contrastive explanations via causal abstraction [7] on smaller open-weight models to establish mechanistic grounding.

REFERENCES

[1] Chirag Agarwal, Eshika Saxena, Satyapriya Krishna, Martin Pawelczyk, Nari Johnson, Isha Puri, Marinka Zitnik, and Himabindu Lakkaraju. 2022. OpenXAI: Towards a Transparent Evaluation of Model Explanations. *Advances in Neural Information Processing Systems* 35 (2022).

[2] David Alvarez-Melis and Tommi Jaakkola. 2018. Towards Robust Interpretability with Self-Explaining Neural Networks. *Advances in Neural Information Processing Systems* 31 (2018).

[3] Arthur Conmy, Augustine N Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards Automated Circuit Discovery for Mechanistic Interpretability. *Advances in Neural Information Processing Systems* 36 (2023).

[4] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. 2017. CARLA: An Open Urban Driving Simulator. *Proceedings of the 1st Annual Conference on Robot Learning* (2017), 1–16.

[5] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. 2022. Toy Models of Superposition. *arXiv preprint arXiv:2209.10652* (2022).

[6] Mohamed Amine Ferrag, Othmane Friha, Burak Kantarci, Norbert Tihanyi, Lucas Cordeiro, Merouane Debbah, et al. 2026. AgentDrive: An Open Benchmark Dataset for Agentic AI Reasoning with LLM-Generated Scenarios in Autonomous Systems. *arXiv preprint arXiv:2601.16964* (2026).

[7] Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal Abstractions of Neural Networks. In *Advances in Neural Information Processing Systems*, Vol. 34. 9574–9586.

[8] Jiahui Guan et al. 2025. World Models for Autonomous Driving: An In-Depth Survey. *arXiv preprint arXiv:2512.18832* (2025).

[9] Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, and Zhiting Hu. 2023. Reasoning with Language Model is Planning with World Model. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (2023), 8154–8173.

[10] Alon Jacovi and Yoav Goldberg. 2020. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020), 4198–4205.

[11] Jianhua Lin. 1991. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory* 37, 1 (1991), 145–151.

[12] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems* 30 (2017).

[13] Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2024. Towards Faithful Chain-of-Thought Reasoning. *Proceedings of the 13th International Joint Conference on Natural Language Processing* (2024).

[14] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing* (2019), 3982–3992.

[15] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016), 1135–1144.

[16] Alexis Ross, Matthew E Peters, and Sebastian Ruder. 2021. Explaining by Removing: A Unified Framework for Model Explanation. *Journal of Machine Learning Research* 22, 209 (2021), 1–90.

[17] Andrea Saltelli. 2002. Making best use of model evaluations to compute sensitivity indices. *Computer Physics Communications* 145, 2 (2002), 280–297.

[18] Sahil Verma, John Dickerson, and Keegan Hines. 2020. Counterfactual Explanations for Machine Learning: A Review. *arXiv preprint arXiv:2010.10596* (2020).

[19] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837.