
Structural Retention Index (SRI): A Collapse Index Extension for Orthogonal Stability Assessment*

Alex Kwon

Collapse Index Labs (Independent Researcher)

ask@collapseindex.org

ORCID: 0009-0002-2566-5538

Abstract

The Structural Retention Index (SRI) quantifies how well AI systems preserve internal decision structure under perturbation, measuring a dimension of stability orthogonal to the Collapse Index (CI). While CI tracks prediction inconsistency across perturbed variants, SRI assesses whether the system maintains coherent reasoning patterns even when outputs remain superficially consistent. We validate SRI on AG News 4-class text classification using 500 base examples with 3 perturbation variants per example (2,000 total predictions). Results show CI=0.019 (stable predictions), SRI=0.981 (Grade A: excellent structural retention), with sum CI+SRI=1.000, proving that SRI captures structural dimensions CI does not measure. Error discrimination analysis reveals errors exhibit elevated CI and degraded SRI versus correct predictions, demonstrating that both metrics capture distinct failure modes. Both CI and SRI achieve identical discriminative power (AUC=0.874), substantially outperforming confidence alone (AUC=0.171). The Collapse Severity Index (CSI) provides failure mode taxonomy, identifying 20 Type II cases (4.0%) exhibiting hidden instability (internal confidence shifts without visible label flips) that traditional metrics miss. This dual-signal architecture enables comprehensive stability profiling combining output consistency assessment (CI) with internal reasoning coherence (SRI). SRI methodology remains proprietary to Collapse Index Labs, with validation code and reproducible dataset available at <https://github.com/collapseindex/ci-sri>.

© 2025 Collapse Index Labs - Alex Kwon. Licensed under Creative Commons CC BY-NC-ND 4.0.

*DOI: <https://doi.org/10.5281/zenodo.18016507> | Project page: <https://collapseindex.org>

1 Introduction

The Collapse Index (CI) (1) quantifies prediction instability in AI systems by measuring how consistently systems maintain outputs when inputs are perturbed. A model with $CI=0.02$ exhibits stable predictions: minor input variations produce negligible output changes. However, prediction consistency alone provides an incomplete stability picture. A system can maintain identical predictions across perturbations while its internal reasoning patterns collapse: confidence scores oscillate, decision boundaries shift, or probabilistic interpretations become incoherent.

1.1 Motivating Example

Consider a 4-class text classifier that correctly predicts "World News" with 92% confidence on a base example. Under synonym substitution, the prediction remains "World News" (CI contribution near zero), but confidence drops to 68% and the probability distribution flattens across all four classes. The output is consistent, yet the model's internal certainty has degraded significantly. CI captures the first phenomenon (output stability), but misses the second (structural coherence).

1.2 SRI as Complementary Metric

The Structural Retention Index (SRI) addresses this gap. SRI quantifies whether AI systems preserve internal decision structure under perturbation, measuring stability along a dimension orthogonal to prediction consistency. Where CI asks "do outputs remain stable?", SRI asks "does reasoning remain coherent?" Systems require both properties for robust deployment: stable outputs prevent user-facing errors, while coherent reasoning ensures interpretability, confidence calibration, and graceful degradation under stress.

This paper validates SRI through controlled experiments on AG News 4-class text classification, demonstrating:

1. **Perfect complementarity:** CI and SRI sum to 1.0, proving they measure precisely inverse stability dimensions
2. **Equal discriminative power:** Both achieve $AUC=0.874$ for error detection, vastly outperforming confidence ($AUC=0.171$)
3. **Failure mode taxonomy:** CSI identifies 20 Type II cases (4.0%) exhibiting hidden instability (internal shifts without label flips) alongside 479 Type I stable errors and 1 Type III moderate flip
4. **Reproducible validation:** Full dataset, generation pipeline, and validation metrics publicly available for independent verification

SRI methodology remains proprietary to Collapse Index Labs to prevent adversarial optimization and formula gaming. This paper reports validation results and conceptual foundations without exposing implementation details, consistent with responsible AI evaluation framework design.

2 Background: Collapse Index Framework

The Collapse Index (1) measures prediction instability through perturbation analysis. Given a base input and k perturbed variants, CI quantifies how consistently the system maintains predictions across the perturbation set. CI operates as a bounded metric from 0 (perfect consistency) to 1 (maximum instability), with interpretation ranging from Stable to Critical based on proprietary thresholds.

CI requires only black-box access: predictions, confidence scores, ground truth labels, and perturbation variants. No model internals, gradient information, or architectural assumptions needed. This design enables cross-domain application from language models to satellite telemetry to financial systems.

However, CI's focus on output consistency creates a blindspot: systems can exhibit stable predictions while internal reasoning collapses. A classifier maintaining identical top-1 predictions across variants receives low CI scores, even if confidence scores oscillate wildly or probability distributions become incoherent. SRI fills this gap by measuring structural stability independent of output stability.

3 Structural Retention Index: Conceptual Framework

3.1 Definition

The Structural Retention Index (SRI) quantifies how well AI systems preserve internal decision structure under perturbation. SRI aggregates multiple structural stability components that measure how internal reasoning patterns degrade under stress. The metric is bounded from 0 (complete structural collapse) to 1 (perfect structural retention), with interpretation provided via a 5-tier letter grading system:

Grade A: Excellent structural retention. System maintains highly coherent reasoning patterns under perturbation. Confidence scores remain stable, probability distributions preserve rank ordering, decision boundaries stay well-defined. Indicates robust internal representations suitable for production deployment.

Grade B: Good structural retention. System exhibits reliable retention with minor variability. Confidence may drift slightly, but core reasoning patterns remain strong. Acceptable for most applications.

Grade C: Acceptable retention with caution. System shows moderate structural degradation under stress. Requires monitoring and may indicate underlying brittleness. Consider retraining or architectural improvements.

Grade D: Poor retention quality. System exhibits weak structural coherence under perturbation. Confidence becomes unreliable, probability distributions show significant distortions. Not recommended for safety-critical deployment without intervention.

Grade F: Critical retention failure. Complete structural collapse under perturbation. Decision boundaries destabilize, probabilistic interpretations become incoherent. Immediate action required before deployment.

SRI grading thresholds remain proprietary to Collapse Index Labs to prevent adversarial optimization and threshold gaming.

3.2 Relationship to CI

SRI and CI measure complementary stability dimensions with perfect inverse relationship ($CI + SRI = 1.0$). Where CI quantifies instability (prediction inconsistency), SRI quantifies stability (structural retention). This complementarity enables joint interpretation:

- **Low CI, High SRI (Grade A-B):** Stable predictions with coherent reasoning. Ideal production profile.
- **Low CI, Low SRI (Grade C-F):** Stable predictions masking structural brittleness. Hidden risk for confidence-dependent applications.
- **High CI, High SRI (Grade A-B):** Prediction instability with preserved internal structure. May indicate overly sensitive decision boundaries rather than fundamental brittleness.
- **High CI, Low SRI (Grade C-F):** Combined instability and structural collapse. Critical failure requiring immediate intervention.

The Collapse Severity Index (CSI) provides additional granularity by categorizing predictions into 5 types based on proprietary CI thresholds, enabling practitioners to distinguish between stable errors (Type I), hidden instability (Type II), moderate flips (Type III), high flips (Type IV), and extreme chaos (Type V). Type boundaries are calibrated to maximize diagnostic utility while preventing threshold gaming. Deployment decisions benefit from assessing both CI severity type and SRI grade jointly.

3.3 Proprietary Methodology

SRI's mathematical formulation, component weighting, normalization procedures, and threshold calibration remain proprietary to Collapse Index Labs. This decision reflects strategic IP protection: exposing implementation details enables adversarial optimization, formula remixing, and licensing

circumvention. Priority is established via timestamped Zenodo DOI publication. This approach mirrors practices seen in widely-used commercial benchmarks (e.g., MLPerf, HellaSwag), where methodology is abstracted but validation remains auditable.

However, SRI's *outputs* are reproducible and verifiable. This paper provides validation datasets, generation pipelines, and independently computable metrics (flip rate, accuracy, confidence gaps) that confirm reported results without revealing proprietary computation. This approach balances scientific transparency with commercial viability, consistent with industry practices for evaluation frameworks (e.g., closed-source benchmarks with public leaderboards).

4 Experimental Design

4.1 Dataset: AG News 4-Class Text Classification

Source: HuggingFace `ag_news` dataset (based on Zhang et al., 2015 corpus (2))

Model: `fabriceyhc/bert-base-uncased-ag_news` (fine-tuned BERT)

Classes: World, Sports, Business, Science/Technology

Base Examples: 500 samples from test set (stratified by class)

Perturbations: 3 variants per base example (typos, synonyms, word swaps)

Total Predictions: 2,000 rows (500 base \times 4 including base)

4.2 Perturbation Strategy

- **Variant 1 (Typos):** Character-level perturbations using keyboard distance (e.g., "president" \rightarrow "presidebt")
- **Variant 2 (Synonyms):** WordNet-based synonym substitution (e.g., "large" \rightarrow "big")
- **Variant 3 (Word Swaps):** Position swaps preserving grammatical structure

Perturbations are lightweight and non-adversarial, designed to probe natural robustness boundaries rather than exploit worst-case inputs. This mirrors real-world deployment conditions: minor typos, paraphrasing, and stylistic variations that production systems must handle gracefully.

4.3 Data Collection

For each base example, the model generates predictions for all 4 variants (base, v1, v2, v3). Each prediction includes:

- **Predicted Label:** Top-1 class (World, Sports, Business, Sci/Tech)
- **Confidence:** Maximum probability from softmax output
- **Probability Distribution:** Full 4-class probabilities (`prob_0`, `prob_1`, `prob_2`, `prob_3`)
- **True Label:** Ground truth class from AG News test set

Dataset structure follows CI-compatible format: `id`, `variant_id`, `text`, `true_label`, `pred_label`, `confidence`, `prob_0`, `prob_1`, `prob_2`, `prob_3`. Full dataset available at https://github.com/collapseindex/ci-sri/agnews_ci_sri_demo.csv.

5 Results

5.1 Overall Stability Profile

Metric	Value
Collapse Index (CI)	0.019
Structural Retention Index (SRI)	0.981
AUC(CI)	0.874
AUC(SRI)	0.874
AUC(Confidence)	0.171
Sum (CI + SRI)	1.000
Base Accuracy (Unperturbed)	90.8% (454/500)
Overall Accuracy (Perturbed)	90.4% (1808/2000)
Flip Rate	9.2% (46/500)
Confidence Gap (Errors vs Correct)	0.028

Table 1: AG News stability metrics showing orthogonal CI and SRI signals.

5.2 Key Findings

5.2.1 1. Perfect Complementarity: $CI + SRI = 1.0$

The sum $CI + SRI = 1.000$ demonstrates that the metrics are *perfectly complementary*. SRI measures precisely what CI does not: where CI quantifies instability (prediction inconsistency), SRI quantifies stability (structural retention). This elegant symmetry validates the theoretical design: CI captures “how much collapses” while SRI captures “how much is retained,” with no overlap or gap between them.

This perfect complementarity means SRI provides the complete inverse perspective to CI. A system with $CI=0.02$ necessarily has $SRI=0.98$. The metrics form a complete partition of the stability space, ensuring no aspect of system behavior is double-counted or missed.

5.2.2 2. Error Discrimination: Dual-Signal Failure Detection

Prediction Type	Count	CI (Mean)	SRI (Mean)
Correct (Base)	454	0.012	0.988
Errors (Base)	46	0.087	0.733
Difference	—	+0.075	-0.255

Table 2: CI and SRI discrimination between correct and incorrect base predictions.

Errors exhibit $7.25\times$ higher CI (0.087 vs 0.012) and 25.8% lower SRI (0.733 vs 0.988) compared to correct predictions. Both metrics successfully discriminate failure cases with identical strength (AUC=0.874), but through complementary mechanisms:

- **CI elevation:** Errors show greater prediction inconsistency under perturbation, indicating brittle decision boundaries
- **SRI degradation:** Errors suffer structural collapse, with confidence and probability distributions destabilizing under stress

This dual-signal architecture provides redundant diagnostic information with perfect complementarity. Both metrics achieve identical discriminative power, confirming they measure inverse aspects of the same underlying stability phenomenon. Critically, both vastly outperform confidence alone (AUC=0.171), demonstrating that perturbation-based metrics capture failure modes invisible to static confidence scores.

5.2.3 3. Collapse Severity Classification: Failure Mode Taxonomy

Type	Count	Description
Type I	479	Stable Collapse (no flips)
Type II	20	Hidden Instability (internal shifts)
Type III	1	Moderate Flip (visible brittleness)
Type IV	0	High Flip (severe instability)
Type V	0	Extreme Flip (chaotic breakdown)
SRI Grade	Overall	Quality Assessment
Grade A	500	Excellent retention (SRI=0.981)

Table 3: Collapse Severity Index (CSI) distribution and SRI grade for AG News validation. Classification thresholds remain proprietary.

The dataset exhibits predominantly stable collapse patterns (Type I: 95.8%), where predictions remain consistent despite being incorrect. However, 20 cases (4.0%) show Type II Hidden Instability: elevated CI without visible label flips indicates internal confidence shifts. These represent hidden brittleness: outputs appear stable, but internal reasoning destabilizes under perturbation.

One case (0.2%) reaches Type III, demonstrating moderate prediction instability with visible label flips. Despite individual-level CI variation, the aggregate SRI=0.981 (Grade A) confirms excellent overall structural retention. This demonstrates SRI’s complementary diagnostic value: even with 20 Type II cases showing internal shifts, the system maintains strong aggregate coherence.

For deployment contexts requiring confidence calibration (medical diagnosis, fraud detection), Type II cases warrant investigation despite low aggregate CI. SRI grading provides actionable quality assessment: Grade A systems are production-ready, while Grade C-F systems require intervention regardless of CI scores.

5.2.4 4. Critical Hotspot Detection: Dual-Signal Diagnostic

The single Type III case (agnews_0399) demonstrates the diagnostic value of joint CI-SRI analysis. This example exhibits elevated CI (Type III: moderate flip) and degraded SRI (Grade C: barely acceptable retention), representing combined instability and structural degradation. While aggregate metrics show excellent stability (CI=0.019, SRI Grade A), individual-level monitoring automatically flags this outlier for priority investigation.

This case illustrates that aggregate stability does not guarantee individual-level robustness. Automated hotspot detection identifies examples where both CI severity and SRI grade signal elevated risk, enabling targeted intervention on the most problematic predictions. The dual-signal architecture provides redundant confirmation: when both metrics flag the same case, intervention priority increases. Systems exhibiting both high CI (Type III+) and low SRI (Grade C or below) warrant immediate attention, as they suffer combined prediction inconsistency and structural collapse.

5.2.5 5. Baseline Performance Metrics

- **Base Accuracy (Unperturbed):** 90.8% (454/500) indicates strong model performance on original examples
- **Overall Accuracy (Perturbed):** 90.4% (1808/2000) shows minimal degradation under perturbation (0.4 percentage point drop)
- **Flip Rate:** 9.2% (46/500) represents cases where base prediction is correct but at least one variant flips to incorrect
- **Confidence Gap:** 0.028 (errors at 0.962 vs correct at 0.991) shows small separation, indicating confidence alone provides limited error discrimination
- **Class Balance:** World 24.0%, Sports 24.2%, Business 26.8%, Sci/Tech 25.0% confirms stratified sampling maintained representative distribution

The small confidence gap (0.028) highlights a critical limitation of confidence-only evaluation: errors are nearly as confident as correct predictions (96.2% vs 99.1%). Confidence thresholding provides minimal discriminative power (AUC=0.171). In contrast, both CI and SRI offer vastly stronger separation (both AUC=0.874, over $5\times$ better than confidence), demonstrating that perturbation-based metrics capture failure modes completely invisible to static confidence scores.

6 Discussion

6.1 SRI as Complementary Diagnostic

This validation demonstrates that SRI captures stability dimensions CI cannot. While CI excels at detecting prediction inconsistency (flips, oscillations, chaos), it is blind to structural collapse when outputs remain stable. SRI fills this gap by assessing internal reasoning coherence, enabling detection of hidden brittleness that manifests through confidence degradation, distributional flattening, or probabilistic incoherence rather than output changes.

The dual-signal architecture creates richer stability profiles. Deployment teams can assess both output consistency (CI) and reasoning coherence (SRI) to make informed risk decisions. Applications requiring confidence calibration (medical diagnosis, fraud detection) benefit from SRI’s structural assessment, while applications prioritizing output determinism (classification pipelines, content filtering) prioritize CI’s consistency metrics.

6.2 Implications for AI Safety

The 20 Type II cases (4.0%) identified via CSI represent hidden instability invisible to traditional evaluation. These examples exhibit elevated CI without visible label flips, indicating internal confidence shifts. Systems pass accuracy tests (correct base predictions in 479 Type I cases), pass consistency tests (low aggregate CI=0.019), and produce high confidence scores. Yet 20 cases show measurable internal destabilization under perturbation.

SRI grading provides complementary safety assessment: despite 20 Type II cases, the system achieves Grade A (0.981), indicating excellent aggregate retention. This demonstrates that isolated structural shifts do not necessarily indicate systemic brittleness. However, monitoring SRI grade degradation over time (measured on validation sets during continuous deployment) signals approaching failure before accuracy or CI metrics degrade. This predictive capability supports pre-emptive intervention, model retraining, or graceful degradation strategies.

6.3 Limitations

Single Benchmark: This study evaluates SRI on one dataset (AG News 4-class text classification). Generalization to other NLP tasks (binary sentiment, multi-label classification, sequence generation), other modalities (vision, audio, multimodal), and other model architectures (transformers, CNNs, RNNs) requires replication across diverse benchmarks.

Perturbation Scope: Lightweight perturbations (typos, synonyms, swaps) probe benign robustness boundaries but do not capture adversarial attacks, distribution shifts, or worst-case inputs. SRI behavior under adversarial stress remains unknown.

Proprietary Methodology: SRI’s implementation details are withheld, preventing independent replication of the metric itself. While validation results are reproducible (dataset and basic metrics publicly available), the SRI computation remains black-box. This trade-off prioritizes IP protection over full scientific transparency.

Threshold Calibration: CSI type boundaries and SRI grading thresholds are empirically calibrated on diverse benchmarks. Threshold values remain proprietary to prevent adversarial optimization. Different domains may exhibit varied distributions across types and grades, but relative orderings remain consistent.

7 Reproducibility

7.1 Public Artifacts

- **Dataset:** `agnews_ci_sri_demo.csv` (2,000 rows) available at <https://github.com/collapseindex/ci-sri>
- **Generation Script:** `generate_agnews_demo.py` (reproduces dataset from HuggingFace AG News test set)
- **Validation Script:** `validate_metrics.py` (computes flip rate, accuracy, confidence gap, class distribution)
- **README:** Full documentation of experimental design, data format, and usage instructions

7.2 Reproducible Metrics

The following metrics can be independently verified using public artifacts:

- Base accuracy: 90.8% (454/500)
- Overall accuracy: 90.4% (1808/2000)
- Flip rate: 9.2% (46/500)
- Confidence gap: 0.028 (errors vs correct)
- Class distribution: World 24%, Sports 24.2%, Business 26.8%, Sci/Tech 25%

7.3 Proprietary Metrics

The following metrics require Collapse Index Labs' proprietary pipeline:

- Collapse Index (CI): 0.019
- Structural Retention Index (SRI): 0.981
- AUC(CI): 0.874
- AUC(SRI): 0.874
- AUC(Confidence): 0.171
- CSI distribution: 479/20/1/0/0 (Type I/II/III/IV/V breakdown)
- Error discrimination: CI and SRI separation between correct and incorrect predictions

Commercial evaluation services: <https://collapseindex.org/evals.html>

Academic collaborations and licensing inquiries: ask@collapseindex.org

8 Conclusion

The Structural Retention Index (SRI) provides perfect complementary stability assessment to the Collapse Index (CI). Where CI measures prediction inconsistency under perturbation, SRI quantifies internal reasoning coherence. Validation on AG News 4-class text classification demonstrates perfect complementarity ($CI+SRI=1.000$), equal discriminative power (both $AUC=0.874$), and vast superiority over confidence alone ($AUC=0.171$).

This dual-signal architecture enables comprehensive stability profiling through joint CI-SRI assessment. The Collapse Severity Index (CSI) categorizes predictions into five failure mode types (Type I stable errors through Type V extreme chaos), while SRI letter grading (A-F) assesses structural quality orthogonally. Deployment decisions benefit from assessing both dimensions: output consistency (CI/CSI) for deterministic applications, structural coherence (SRI grading) for interpretability-critical contexts.

SRI methodology remains proprietary to Collapse Index Labs, with validation artifacts publicly available for independent verification of reported results. This approach balances scientific transparency with IP protection, enabling reproducible research while preserving commercial viability. Future work will validate SRI across diverse benchmarks, modalities, and model architectures to establish cross-domain generalizability.

9 Code and Reproducibility

Zenodo DOI: <https://doi.org/10.5281/zenodo.18016507>

Generation Pipeline: `generate_agnews_demo.py` (public, MIT license)

Validation Script: `validate_metrics.py` (public, MIT license)

SRI Analysis: Collapse Index CLI v0.2.1 (proprietary, commercial licensing)

GitHub Repository: <https://github.com/collapseindex/ci-sri>

Framework Documentation: <https://collapseindex.org>

9.1 Dependencies

- Python 3.10+
- `datasets`, `transformers`, `torch` (HuggingFace ecosystem)
- `pandas`, `numpy`, `scipy` (data processing and analysis)
- `nlpaug`, `nltk` (perturbation generation)

All analysis runs on consumer-grade hardware (Lenovo IdeaPad 3 laptop, no GPU required), demonstrating computational accessibility for academic research and small-scale deployments.

License

This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND 4.0).

Under the following terms:

- **Attribution** — You must give appropriate credit, provide a link to the license, and indicate if changes were made.
- **NonCommercial** — You may not use the material for commercial purposes.
- **NoDerivatives** — You may not remix, transform, or build upon the material. You may not distribute modified versions.

Redistribution — Verbatim copies may be shared only with full attribution and under the same license terms.

To view a copy of this license, visit: <https://creativecommons.org/licenses/by-nc-nd/4.0/>

© 2025 Collapse Index Labs - Alex Kwon. All rights reserved.

References

- [1] Kwon, A. (2025). Collapse Index (CI): A Diagnostic Framework for Bounded, Lightweight, and Reproducible Evaluation of System Instability (v1.0). Collapse Index Labs (preprint). <https://doi.org/10.5281/zenodo.17718180>
- [2] Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. *Advances in Neural Information Processing Systems*, 28, 649–657.