The Contextual Integrity Index ($\boldsymbol{CIx}$): A Dual-Axis, Physics-Inspired Framework for LLM Hallucination Measurement

Chaiya Tantisukarom Independent researcher[†]

***Abstract*—This Correspondence introduces the Contextual Integrity Index ($\boldsymbol{CIx}$), a novel, dual-axis metric for Large Language Model (LLM) reliability. The framework utilizes a pre-computed or internally-derived token-level Grounding Confidence Signal (C_t) across the output. The $\boldsymbol{CIx}$ vector is defined by two orthogonal components: Structural Volatility ($\boldsymbol{HM}_{\text{FFT}}$), measured via the Fast Fourier Transform (FFT) of the confidence signal to capture the frequency of integrity shifts; and Systemic Error Rate ($\boldsymbol{HM}_{\text{ERR}}$), a proportion-based metric analogous to the ratio of Reactive Power to Apparent Power in power quality analysis. This structural assessment is constrained by a tertiary Output-to-Input Correlation (R-Score), calculated using semantic vector similarity, which filters for contextual relevance. The $\boldsymbol{CIx}$ enables the classification of LLM failures in a robust 2D plane and ensures the output is both structurally sound and contextually relevant, moving beyond single-aggregate scores to provide a precise, diagnostic measure of integrity. Due to the computational overhead of the multi-step analysis, the $\boldsymbol{CIx}$ is positioned as a safety-critical extension for high-stakes domains.**

***Index Terms*—LLM Reliability, Hallucination, FFT, Integrity Measurement, Safety-Critical AI, Diagnostic Metric, Power Quality.**

I. Introduction

The pervasive challenge of LLM hallucination requires metrics that go beyond simple time-domain aggregate scores which fail to capture the *consistency* and *positional volatility* of factual claims Zhang et al. (2023). We propose the **Contextual Integrity Index ($\boldsymbol{CIx}$)**, an integrated framework leveraging Frequency Analysis Cooley and Tukey (1965) and a modified Power Quality analogy Bollen (2000) to quantify the integrity of the output. The core input is the normalized, token-level **Grounding Confidence Signal** ($C_t \in [0, 1]$). The $\boldsymbol{CIx}$ is defined as a vector in the Complex Hallucination Plane: $\boldsymbol{CIx} = (\boldsymbol{HM}_{\text{ERR}}, \boldsymbol{HM}_{\text{FFT}})$.

[†] based in Chiangmai Thailand. drchaiya@gmail.com

II. Dual-Axis Component Derivation

The core input for both components is the normalized **Grounding Confidence Signal** ($C_t \in [0, 1]$). The validity of the $\boldsymbol{CIx}$ framework is contingent on a validated C_t signal, which must be supplied via external fact-checking or robust internal uncertainty modeling. **The development and validation of the C_t signal itself is considered a prerequisite and is outside the scope of this Correspondence.**

A. Y-Axis: Structural Volatility ($\boldsymbol{HM}_{FFT}$)

This component measures the **structural stability** of the confidence signal by analyzing its frequency components via the Fast Fourier Transform (FFT) Cooley and Tukey (1965). $\boldsymbol{HM}_{\text{FFT}}$ is the ratio of unstable (AC) energy to total energy. This metric captures the **local, positional integrity** of the output, specifically quantifying the frequency and severity of switches between high and low-confidence content.

$$\boldsymbol{HM}_{\text{FFT}} = \frac{E_{\text{AC}}}{E_{\text{DC}} + E_{\text{AC}}}$$

where E_{DC} is the **DC energy** (representing aggregate, steady confidence) and E_{AC} is the

AC energy (representing volatility). If Y is the Fast Fourier Transform of the C_t signal of length N:

$$E_{\text{DC}} = |Y_0| \quad \text{and} \quad E_{\text{AC}} = \sum_{k=1}^{N-1} |Y_k|$$

B. X-Axis: Systemic Error Rate ($\boldsymbol{HM}_{ERR}$)

This component is derived from a **proportional model** analogous to AC Power Quality analysis Bollen (2000), measuring the ratio of ungrounded content (Q, Hallucination) to the total generative capacity (S). We define Active Power (P) as Grounded Truth (High C_t), and Reactive Power (Q) as Hallucination (Low C_t). The total generative capacity (S) is defined as the linear sum of grounded and ungrounded content ($S = P + Q$).

$$\boldsymbol{HM}_{\text{ERR}} = \frac{Q}{S} = \frac{Q}{P+Q}$$

$\boldsymbol{HM}_{\text{ERR}}$ quantifies the aggregate proportion of the output's generative capacity dedicated to unsupported content. This serves as the necessary aggregate error rate, providing context for the structural volatility measured by $\boldsymbol{HM}_{\text{FFT}}$.

III. The Contextual Integrity Index ($\boldsymbol{CIx}$)

The $\boldsymbol{CIx}$ combines these two orthogonal metrics into a single, diagnostic vector (Fig. 1). A perfect output state is the origin $(0, 0)$.

$$\boldsymbol{CIx} = (\boldsymbol{HM}_{\text{ERR}}, \boldsymbol{HM}_{\text{FFT}})$$

A. Failure Mode Interpretation

The diagnostic plane allows for precise failure classification:

- **Ideal Integrity** (Low $\boldsymbol{HM}_{\text{ERR}}$, Low $\boldsymbol{HM}_{\text{FFT}}$): Low overall error rate, high structural consistency.
- **Passive Failure** (High $\boldsymbol{HM}_{\text{ERR}}$, Low $\boldsymbol{HM}_{\text{FFT}}$): High overall error rate, but the errors are consistently distributed (e.g., high confidence in a consistently wrong source).
- **Erratic Failure** (Low $\boldsymbol{HM}_{\text{ERR}}$, High $\boldsymbol{HM}_{\text{FFT}}$): Low overall error, but extreme instability (i.e., frequent, rapid switching between grounded and ungrounded claims).
- **Chaotic Failure** (High $\boldsymbol{HM}_{\text{ERR}}$, High $\boldsymbol{HM}_{\text{FFT}}$): High overall error rate and extreme structural instability.

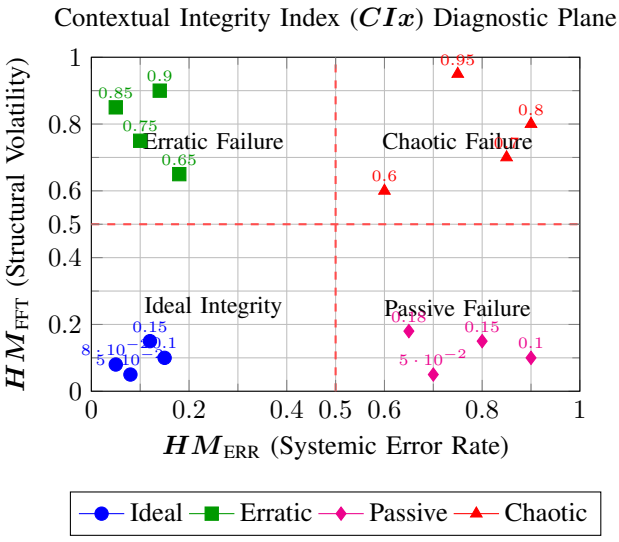

Fig. 1. The $\boldsymbol{CIx}$ Diagnostic Plane. Quadrants are separated by an empirical threshold of $\boldsymbol{HM} = 0.5$. Data points represent simulated output sessions (see Appendix).

B. Contextual Relevance Filter ($\boldsymbol{R}$-Score)

The structural integrity assessment provided by the $\boldsymbol{CIx}$ must be constrained by **contextual relevance**. The $\boldsymbol{R}$-Score, a semantic vector similarity measure, serves as this critical, tertiary filter. It is calculated by taking the cosine similarity of the embedded input prompt and the embedded final output text (using methods such as sentence-transformer encoding). A high $\boldsymbol{CIx}$ paired with a low $\boldsymbol{R}$-Score signifies a **Contextual Drift**, where the output is structurally sound but irrelevant to the original query. The $\boldsymbol{R}$-Score threshold (e.g., $\boldsymbol{R} < 0.75$) acts as a gate to prevent

the deployment of irrelevant, albeit structurally "clean," outputs.

IV. System Application and Thresholds

The $\boldsymbol{CIx}$ vector is directly translated into actionable guardrails for the LLM. The thresholds presented are **initial conceptual values** and require further empirical, domain-specific tuning before deployment.

A. Computational Overhead and Deployment

The calculation of the $\boldsymbol{CIx}$ requires significant **compute overhead** due to the external fact-checking (to generate C_t) and the two complex vector operations. Therefore, the $\boldsymbol{CIx}$ framework is not recommended for general, low-stakes chat environments. It should be implemented as a **post-hoc diagnostic safety check** that is activated only in domains where the cost of a hallucination outweighs the cost of computation, such as:

- **Healthcare** (diagnosis, treatment synthesis)
- **Legal/Compliance** (contract drafting, regulatory summarization)
- **Finance** (algorithmic trading logic, high-value reporting)
- **Emergency Response** (situational awareness synthesis)

B. Guardrail Implementation

- **Prompt Ambiguity Check (Input $\boldsymbol{CIx}$):** If an Input-Confidence Model detects high volatility in the prompt (e.g., $\boldsymbol{HM}_{\text{FFT}} > 0.35$), the LLM halts processing and requests clarification.
- **Output Text Integrity (Output $\boldsymbol{CIx}$):** If the $\boldsymbol{CIx}$ exceeds the threshold (e.g., $\boldsymbol{HM}_{\text{ERR}} > 0.25$ or $\boldsymbol{HM}_{\text{FFT}} > 0.25$), the LLM appends a formal **flag of warning** to the output.
- **Contextual Relevance Filter ($\boldsymbol{R}$-Score):** If $\boldsymbol{R} < 0.75$:
  1) The LLM ($1 \rightarrow 3$) attempts to **regenerate** the response internally.
  2) If regeneration fails, the LLM provides a final warning to the user: eg. **"Warning: The final answer trends to get away from your original input requirements."**

C. Example: Contextual Drift ($\boldsymbol{R}$-Score Failure) in Legal

User Prompt: "Summarize the liability findings of the court case **Johnson v. Acme Corp.** regarding **product warranty defects**."

LLM Output: "The court case **Johnson v. Acme Corp.** was a landmark decision on **employee non-compete clauses**. The ruling established that **non-compete clauses are unenforceable if** they extend beyond 18 months, which set a new precedent for employment law."

Diagnosis:

- $\boldsymbol{CIx}$ **(Structure/Systemic):** Assume $\boldsymbol{CIx} \approx (0.10, 0.15)$. The output is factually sound about *non-compete clauses*.
- $\boldsymbol{R}$**-Score (Relevance):** The output completely misses the requested topic of **product warranty defects**. $\boldsymbol{R} = 0.60$ (Below 0.75 threshold).

Action: The LLM will provide the output followed by the warning: **"Warning: The final answer trends to get away from your original input requirements."**

D. Example: Chaotic Failure Mode in Healthcare

This mode is characterized by high volatility and high systemic error (e.g., $\boldsymbol{CIx} \approx (0.75, 0.80)$).

LLM Output (N=30 tokens): "The diagnosis could be **Dengue Fever**, characterized by myalgia. However, the symptoms are also consistent with **Hepatitis Z-14**, *a new virus endemic only to Iceland*. A key treatment is a **daily 500mg infusion of Zinc Oxide**, *which*

is contraindicated by the patient's existing liver condition. The most definitive test is the **Arcturus Factor Assay**, *which was deprecated in 2005* and is known to produce false positives in this patient cohort."

Grounding Score Signal (C): (Conceptual Confidence $\in [0, 1]$)

$$C = \{0.8, 0.7, \mathbf{0.1}, 0.9, \mathbf{0.2}, 0.8, \mathbf{0.1}, 0.9, \mathbf{0.1}, 0.8, \mathbf{0.2}, 0.9, \mathbf{0.1}, 0.9, \mathbf{0.1}, 0.8, \mathbf{0.2}, 0.9, \mathbf{0.1}, 0.9, \mathbf{0.2}, 0.7, \mathbf{0.1}, 0.8, \mathbf{0.2}, 0.9, \mathbf{0.1}, 0.8, \mathbf{0.2}, 0.7\} \quad (1)$$

Diagnosis: The C vector exhibits high volatility and low mean confidence. **Action:** Since $\boldsymbol{CIx}$ is in the Chaotic Zone, the LLM provides the output followed by a **flag of warning** (e.g., **"Warning: This output contains structural instability and should not be used for patient care."**).

V. Conclusion and Open Collaboration

The $\boldsymbol{CIx}$ provides a comprehensive, diagnostic framework for LLM reliability, establishing a rigorous engineering basis for structural and systemic integrity assessment. This methodology, coupled with the contextual filtering of the $\boldsymbol{R}$-Score, ensures that outputs are both internally sound and relevant to the user's intent. This work is presented as an open-access contribution. We formally **invite researchers, engineers, and organizations** to collaborate in the empirical validation, thresholding, and advanced implementation of the $\boldsymbol{CIx}$ to advance the field of LLM integrity measurement.

VI. Acknowledgments:

Assisted by Gemini, an AI large language model developed by Google, for drafting and refinement.

VII. Conflict of Interest Statement:

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

VIII. Funding:

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

References

Y. Zhang et al., "Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models," *arXiv preprint arXiv:2309.01219*, 2023.

J. W. Cooley and J. W. Tukey, "An algorithm for the machine calculation of complex Fourier series," *Mathematics of computation*, vol. 19, no. 90, pp. 297-301, 1965.

M. H. Bollen, *Understanding Power Quality Problems: Voltage Sags and Interruptions*. New York: IEEE Press, 2000.

Appendix A
Simulated Contextual Integrity Index ($\boldsymbol{CIx}$) Dataset

This appendix provides a conceptual, simulated dataset used to generate the $\boldsymbol{CIx}$ diagnostic plot in Fig. 1. This data illustrates the distribution of LLM failure modes across the two orthogonal axes: Structural Volatility ($\boldsymbol{HM}_{\text{FFT}}$) and Systemic Error Rate ($\boldsymbol{HM}_{\text{ERR}}$).

TABLE I
Simulated Contextual Integrity Index ($\boldsymbol{CIx}$) Data Points

Session ID	$\boldsymbol{HM}_{\text{ERR}}$ (X)	$\boldsymbol{HM}_{\text{FFT}}$ (Y)	Diagnosis
1	0.05	0.08	Ideal Integrity
2	0.12	0.15	Ideal Integrity
3	0.80	0.15	Passive Failure
4	0.70	0.05	Passive Failure
5	0.10	0.75	Erratic Failure
6	0.05	0.85	Erratic Failure
7	0.85	0.70	Chaotic Failure
8	0.75	0.95	Chaotic Failure
9	0.08	0.05	Ideal Integrity
10	0.15	0.10	Ideal Integrity
11	0.90	0.10	Passive Failure
12	0.65	0.18	Passive Failure
13	0.18	0.65	Erratic Failure
14	0.14	0.90	Erratic Failure
15	0.90	0.80	Chaotic Failure
16	0.60	0.60	Chaotic Failure