Mitigating Large Language Model Hallucinations in Industrial Environments via Asymmetric Deterministic Override Architecture

Yi Zeng

***Abstract*—As Generative Artificial Intelligence deepens its application in the Industrial Internet of Things (IIoT) and power system operations, Large Language Models (LLMs) have emerged as critical components in Decision Support Systems (DSS). However, in safety-critical infrastructure such as power grid control and relay protection, LLMs acting as auxiliary decision sources face a fatal Metacognitive Deficit—the inability to recognize their own cognitive blind spots. This deficit frequently causes probabilistically fabricated hallucinated data (e.g., incorrect register mappings or misleading operational instructions) to propagate down to operational decision streams as valid facts. Consequently, human operators are easily misled into executing catastrophic actions.**

To address this challenge, this paper proposes an Asymmetric Deterministic Override Architecture (ADOA). This architecture logically decouples generative processes from a deterministic causal verification layer. The verification layer relies on a lightweight, expert-verified incremental Tri-Evidence Vault. By integrating BM25 term-frequency retrieval with Dense Vector Retrieval in a hybrid algorithm, the layer determines the causal compliance of LLM outputs within microseconds. Upon hitting a blind spot, the verification layer triggers an *Override Veto* to forcibly correct the recommendations provided to operators; if the blind spot is missed but a physical boundary violation is detected, a *Safety Veto* is triggered to intercept and block the dissemination of misleading information.

Theoretical derivations and case studies demonstrate that, in substation protection and control scenarios demanding ultra-low latency (e.g., the < 15ms standard of IEC 61850), the proposed architecture mathematically guarantees the deterministic interception and hard override of hallucinated instructions. This effectively blocks the propagation of erroneous advice to human operators, providing an insurmountable absolute defense for deploying Generative AI in high-security physical domains.

***Index Terms*—Cyber-Physical Systems (CPS), Large Language Models (LLMs), Metacognitive Deficit, Asymmetric Deterministic Override, Sovereign Data Flywheel, Safety-Critical Control**

I. Introduction

The development of Generative AI has significantly enhanced the efficiency of automated program generation and operational scheduling in smart grids and industrial control systems [1]. However, to ensure system safety in practical industrial control, LLMs typically function as Decision Support Systems (DSS) or intelligent assistants, providing threshold recommendations, register offset configurations, or scheduling script suggestions to human operators. Nevertheless, in power grid substations, distribution networks, and critical infrastructure, any register offset, erroneous constant, or logical flaw resulting from an LLM hallucination—once adopted and issued to control terminals by an operator—translates into irreversible physical damage (e.g., equipment burnout or regional blackouts), leading to catastrophic economic and human losses. In other words, the hazard of AI lies not in direct operational failures, but in its ability to output highly deceptive erroneous information that misleads operators into arriving at incorrect conclusions and making fatal mistakes [2], [3].

Behind the rapid evolution of LLM intelligence, we are facing an unprecedented Cyber-Physical Epistemological Crisis [4]. In scenarios where LLMs assist industrial DSS, their cognitive deficits manifest primarily through three typical disaster-inducing mechanisms. As shown in Fig. 1, we comparatively analyze the process of the LLM hallucination cascade avalanche versus the human expert metacognitive halt across three dimensions: topological layout, inference evolution, and cognitive deficit mechanisms:

- **Inference Channels and Topological Layout**: Fig. 1 is divided by a central dashed line into left and right comparison channels. The left column represents the green-toned *Human Expert* causal inference channel, comprising four top-down connected blocks; the right column represents the red/pink-toned *Probabilistic LLM* autoregressive inference channel, also comprising four top-down connected blocks.
- **Temporal Inference and State Evolution Flow**: Given the identical input of a complex industrial control protocol intent (Step 1: *Input Intent*):
  - When a minor deviation occurs in the first inference step of the left human channel (Step 2: *Minor Error*), human-specific metacognitive modules swiftly make a negative judgment (Step 3: *Metacognition Veto*), proactively halting the process and safely correcting it under the calibration of expert rules (Step 4: *[Safe] Halt & Correct*).
  - When the right LLM channel enters a knowledge blind spot in the first inference step and generates a seemingly plausible initial error (Step 2: *Blindspot Entry*), its lack of self-verification leads it to treat this erroneous code as an established factual premise for subsequent generation (Step 3: *Treat error as facts*). This infinitely cascades and amplifies the error, ultimately outputting code that is syntactically correct but physically fatal (Step 4: *[Fatal] Wrong Output*).
- **Cognitive Limitations and Cascade Failure Mech-**

anisms: The profound causes of this hallucination avalanche mechanism lie in the following three intrinsic disaster-inducing characteristics:

1) **Lack of Metacognition (Self-Verification Capability)**: The foundation of AI is probabilistically predicting the next token; subjectively, it lacks the concepts of 'true" and 'false." When outputting absurd instructions or fake configuration parameters, the AI presents them with 100% certainty, treating toxins as truth and making it exceedingly difficult for operators to discern [5].
2) **Complete Loss of Control Over Blind-Spot Data**: Faced with specific industrial protocols and underlying hardware logic that it has never seen, not been trained on, or been incorrectly trained on, the AI cannot recognize its own ignorance. Unlike human experts, it does not pause to indicate 'I don't know"; instead, it forcibly fabricates perfectly structured physical hallucinations based on its generalized probabilistic space [6].
3) **Cascading Chain of Errors**: AI employs autoregressive generation. If the initial minor deviation is not promptly corrected, the AI adopts this error as a factual premise for subsequent reasoning, sliding further down the wrong path and constructing a logically consistent but physically contradictory false reasoning edifice [6].

Currently, the industry frequently attempts to reduce hallucinations by adding contextual constraints (Prompt Engineering), vertical domain Fine-Tuning, or Reinforcement Learning from Human Feedback (RLHF) [7]. However, fundamentally, the output of neural networks is a probabilistic approximation. When confronting extremely rare industrial faults with the lowest probabilities of occurrence, LLMs—due to a severe lack of underlying training samples—tend to rely on their probabilistic autoregressive mechanisms to forcibly predict and output a seemingly reasonable, high-probability common answer. This over-extrapolation caused by data scarcity not only results in higher logical error rates, but for highly specific anomalies, the model completely fails to make correct judgments and instead misleads operators with extremely high confidence. Therefore, in safety-critical industrial scenarios with zero fault tolerance, soft alignment at the probabilistic layer mathematically cannot provide safety credit endorsements [8].

To tackle the aforementioned challenges, this paper proposes the Laplace Asymmetric Deterministic Override Architecture (ADOA). The topological composition and verification mechanisms of this architecture are illustrated in Fig. 2. Its core logical flow can be strictly decomposed into the following three systemic dimensions:

- **Topological Partitioning and Decoupled Units**: The system's architecture is distributed across three physical zones. The left Execution Domain encompasses operational interfaces (*Operator / User*, *Safety Console*) and the underlying physical grid (*Substation Grid*); the central Decision Domain achieves physical isolation between generation and verification, consisting of a probabilistic generation layer (*Generation Layer* $\mathcal{G}$) in series with a high-priority blocking gateway (*Asymmetric Gate* $\mathcal{V}$); the right Consensus Domain hosts the accumulation of system experience, comprising a local physical truth vault (*Physical Vault* $\mathcal{D}$) and a global federated sharing network (*Federated Platform*).
- **Sequential Instruction and Closed-loop Flow**: System operation is strictly constrained within a 7-step control chain. Step 1 (*Control Request*) initiates the raw intent; Step 2 (*Probabilistic Commands*) generates primary control variables via the LLM; Step 3 (*Causal Override*) triggers core intervention, where the gateway interacts with the truth vault for causal matching and forced override; Step 4 (*Override/Intercept*) returns absolutely safe intercepted or modified instructions to the console; Step 5 (*Zero-Hallucination Command*) executes the final physical action; meanwhile, Steps 6 and 7 (*Audit Log & Consensus*, *Federated Sync*) asynchronously complete federated blockchain auditing of logs and node synchronization of global error-prevention rules.
- **Deterministic Isolation and Consensus Evolution**: The safety completeness of this architecture is built upon two systemic defense lines. First, the hard isolation between probabilistic inference and physical execution. The system refuses to entrust safety guarantees to the intrinsic alignment of probabilistic models; instead, it forcibly converges risks through the unconditional interception (Override Veto) of the gateway layer. Second, the unification of data sovereignty and immune evolution. Through the distributed collaboration of the local truth vault and the federated platform, the system ensures the sovereign privacy of plant and station industrial control data while achieving a collective safety evolutionary synergy of 'single-point error correction, network-wide immunity"

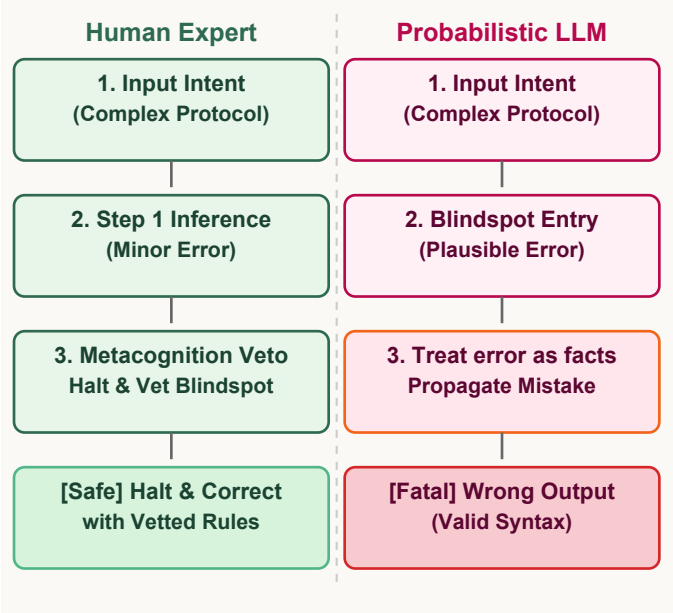

Fig. 1. LLM Hallucination Cascade vs. Human Metacognitive Loop

[9]–[11].

Building upon this topology and workflow, this study distills four core design contributions:

1) **Derivation of the Blind-Spot Hallucination Avalanche Principle:** Formalizes and mathematically derives the hallucination avalanche principle of Softmax-based autoregressive models during blind-spot generation (corresponding to the LLM probabilistic generation and Link 2 in Fig. 2);
2) **Asymmetric Logical Control Topology:** Proposes an asymmetric logical control topology that separates the probabilistic generation layer used for natural language translation from the causal deterministic layer used for safety verification, ensuring safety-compliant interception at extremely low computational costs (corresponding to the decoupled two-layer architecture and Links 2, 4 in Fig. 2);
3) **Tri-Evidence Verification Structure:** Introduces a Tri-Evidence verification structure containing physical waveforms, industry standard clauses, and errata links, guaranteeing the immutability of the system's truth vault (corresponding to the Physical Vault and Link 3 in Fig. 2);
4) **Sovereign Data Flywheel Mechanism:** Designs a Sovereign Data Flywheel mechanism to resolve industrial data isolation dilemmas [9], [10]. While protecting nodal data privacy, it achieves the federated sharing and evolution of safety rules via consensus and asynchronous synchronization (corresponding to the federated platform and the closed-loop Links 6, 7 in Fig. 2).

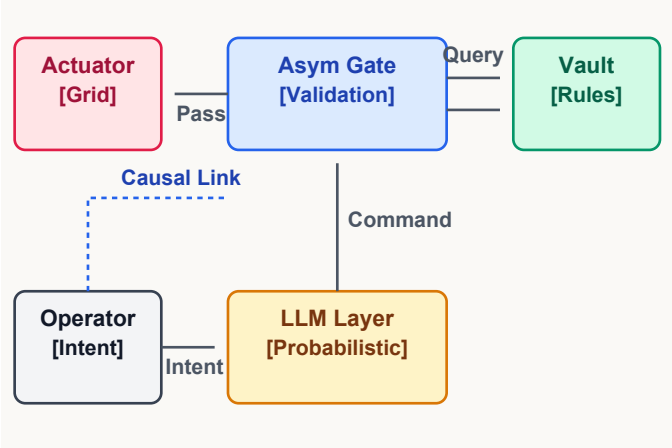

Fig. 2. Laplace Asymmetric Verification Architecture

II. Related Work

A. LLM Hallucination Mitigation Techniques and their Industrial Limitations

Research on mitigating hallucinations in Large Language Models for Decision Support Systems (DSS) has become a focal point in recent years. Mainstream solutions generally fall into three categories:

1) **Offline Calibration via Alignment and Contrastive Learning**: Examples include Reinforcement Learning from Human Feedback (RLHF) and Direct Preference Optimization (DPO) [7]. These methods constrain Softmax probabilities during the model fine-tuning stage by constructing penalizing samples [8]. However, under the long-tail distribution of specific industrial control protocols (e.g., IEC 61850, Modbus), because negative samples cannot be exhausted and the underlying model relies on probabilistic autoregressive sampling, the probabilistic release of logical deviations cannot be entirely eliminated.
2) **Structured Graph Constraints via Knowledge Graphs**: These methods map text representations into ontological knowledge bases [12]. However, they suffer from semantic representation bottlenecks when processing industrial control logic and dynamic register combinations, and dynamic knowledge extraction imposes extreme overhead on microsecond-level operational latency requirements.
3) **Retrieval-Augmented Generation (RAG)**: These techniques introduce large volumes of industry standard documents via "Fat" contexts [13]. Although RAG significantly reduces knowledge-based hallucinations in general semantic tasks, its inference latency (typically milliseconds to seconds) violates the hard real-time constraints (under 15ms) required for dispatching protection and control instructions in smart substations [14]. Furthermore, LLMs still suffer from the "Lost in the Middle" phenomenon within long contexts, failing to provide deterministic safety boundaries.

B. Cyber-Physical System (CPS) Security Defense Mechanisms

In traditional Cyber-Physical Systems (CPS) defense, security strategies primarily focus on boundary and logical credibility assurance [15]. Traditional defenses are mainly categorized into three streams:

1) **Deep Packet Inspection (DPI) for Baseband and Protocols** [16]: Industrial firewalls operating at the network layer can effectively intercept protocol format distortions and unauthorized sources [17]. However, they cannot intercept logically correct but semantically fatal "hallucinated codes" (e.g., mistakenly setting a generator's excitation voltage parameter to ten times its original value), because such instructions comply with all message validation specifications of the communication protocol.
2) **Formal Verification** [18]: Examples include model checking based on temporal logic. While these methods provide strong mathematical guarantees, formal analysis faces the challenge of State-space Explosion under dynamic, complex Industrial IoT network topologies, making real-time, millisecond-level dynamic resolution of decision streams impossible.
3) **Static Rule Compilers and Analyzers**: These perform syntax analysis and symbolic execution on dispatched programs, but lack the capability to flexibly integrate dynamic operator intent, expert prior knowledge bases,

and real-time Tri-Evidence chains. Additionally, security standards for Intelligent Electronic Devices (IED) and SCADA systems, such as IEEE Std 1686, also face bottlenecks in parsing dynamic logical hallucinations [19].

C. Asymmetric Safety Design Principles and Engineering Precedents

In extreme high-trust scenarios such as nuclear power protection, avionics, and rail transit control, there are precedents for defensive designs that decouple the "Generation Layer" from the "Verification Layer." The most representative research is the **Simplex Architecture** proposed by Lui Sha [11]. Its core design principle is "Asymmetric Safety Isolation": The system consists of a "Complex Controller" utilizing advanced but unverified high-performance control algorithms (e.g., probabilistic models, adaptive control), and a "Safety Enforcer Controller" featuring extreme structural simplicity, rigorous formal verification, and SIL 4 safety certification. When the output of the complex controller exceeds preset physical stability boundaries, the safety controller unconditionally revokes its control authority, bringing the system back to a safe and stable state through a hard override (Veto & Override).

The ADOA architecture in this study represents an evolutionary adaptation of the Simplex principle for the era of LLM decision support: It abandons the hope that LLMs (the complex probabilistic layer) will not make mistakes, and instead employs an independent, lightweight expert Causal Layer with an asymmetric topology to execute millisecond-level hard override filtration. This ensures absolute logical safety before "untrusted outputs" are adopted and dispatched by human operators [20].

III. Mathematical Formulation of the Autoregressive LLM Metacognitive Deficit and Cascade Failure Mechanism

A. The Softmax Probability Trap in Blind-Spot Outputs

For an autoregressive Large Language Model based on the Transformer architecture [21], the output vocabulary probability distribution for generating the k-th Token T_k is determined by the Softmax mapping:

$$P(T_k = v \mid T_{<k}) = \frac{\exp(h_k^T W_v)}{\sum_{w \in \mathcal{V}} \exp(h_k^T W_w)} \tag{1}$$

To facilitate cross-disciplinary understanding, we deconstruct this composite formula into three classic theoretical components for physical interpretation:

- **Autoregressive Conditional Probability** $P(T_k = v \mid T_{<k})$: Originating from the Autoregressive Factorization framework in Claude Shannon's Information Theory [21], this represents the theoretical probability of the model predicting the specific token v given all preceding instruction sequences (historical context $T_{<k}$).
- **Bilinear Inner Product Projection** $h_k^T W_v$: Here, $h_k \in R^d$ is the attention representation vector output by the decoder at step k (encapsulating the current industrial context), and $W_v \in R^d$ is the weight vector of the candidate word v in the vocabulary space $\mathcal{V}$. Their dot product derives from the embedding space similarity calculation mechanism in modern Representation Learning, measuring the unnormalized semantic match (i.e., Logit score) between the current contextual state and the candidate action within a high-dimensional continuous space.
- **Softmax Exponential Normalization** $\frac{\exp(x_i)}{\sum_j \exp(x_j)}$: Its mathematical structure is equivalent to the Gibbs-Boltzmann Distribution in statistical physics [22]. If the physical energy state of a candidate action is defined as $E_v = -h_k^T W_v$, the mapping strictly conforms to $P(v) = \frac{\exp(-E_v)}{\sum_w \exp(-E_w)}$. This mechanism forces the exponential scores of all possible vocabulary words to be normalized against the sum over the entire vocabulary space $\mathcal{V}$, ensuring system outputs comply with standard probability axioms (the probability of each candidate action is greater than 0, and the sum strictly equals 1).

We formalize the "cognitive blind spot" of the model from the perspective of Cyber-Physical Epistemology. To achieve analytical tractability for anomaly detection, we assume that the representation vectors h corresponding to normal industrial logic within the model's Supervised Fine-Tuning (SFT) dataset approximately follow a Multivariate Gaussian distribution in the latent space, with mean $\boldsymbol{\mu}_0 \in R^d$ and covariance matrix $\boldsymbol{\Sigma}_0 \in R^{d \times d}$.

When the input prompt falls into an Out-Of-Distribution (OOD) blind spot the model has never seen, the Softmax-based mapping easily exhibits overconfidence when facing OOD samples [23]. Therefore, we must delve into the hidden representation space of the neural network, characterizing the degree of this deviation by calculating the Mahalanobis Distance between the attention vector h_k and the center of the safe industrial logic distribution [24]:

$$d_M(h_k) = \sqrt{(h_k - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}_0^{-1} (h_k - \boldsymbol{\mu}_0)} \tag{2}$$

In extreme blind-spot scenarios, $d_M(h_k) \gg \theta$ (where θ is the normal representation safety confidence threshold bounded by a χ^2 distribution). As revealed by the research of Ethayarajh et al. [25], normal representations in Transformers equipped with correct physical logic comprehension exhibit high anisotropy in latent space, focusing within narrow semantic directional cones. However, in blind spots, because effective underlying physical protocol features cannot be extracted, this anisotropic representation space completely collapses, losing correct semantic directionality. For extremum analysis, this physical degradation process can be modeled as a regression to the global statistical mean superposed with isotropic noise:

$$h_k = \boldsymbol{\mu}_0 + \boldsymbol{\delta}_k \tag{3}$$

where $\boldsymbol{\delta}_k \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$ is an isotropic Gaussian random noise vector devoid of all structured information.

However, because the Softmax in Equation (1) strictly requires the sum of conditional probabilities of its output vector to be 1:

$$\sum_{v \in \mathcal{V}} P(T_k = v \mid T_{<k}) = 1 \tag{4}$$

This deprives traditional LLMs of a mathematical retreat, such as "refusing to answer" or outputting a "safety anomaly (Null)." Substituting the collapsed state of Equation (3) into the bilinear projection to calculate Logit scores yields $z_v = h_k^T W_v = (\boldsymbol{\mu}_0 + \boldsymbol{\delta}_k)^T W_v = \boldsymbol{\mu}_0^T W_v + \boldsymbol{\delta}_k^T W_v$. Since $\boldsymbol{\delta}_k$ is zero-mean white noise, its inner product $\boldsymbol{\delta}_k^T W_v$ degrades to a minor random fluctuation; the output distribution of the system will be entirely dominated by the prior background bias $\boldsymbol{\mu}_0^T W_v$ (i.e., the most frequent and normative global industrial operations memorized by the LLM during the SFT phase). This theoretically explains the critical safety pain point addressed in this paper: when facing rare, unknown industrial faults, the LLM, afflicted by representation collapse, leans toward generating the "common safe operations" with the highest mean in its training set; meanwhile, the normalization property of Softmax still assigns an extremely high local conditional probability to this output. The analysis indicates that such high-confidence misjudgment mechanisms are bound by the model's fundamental architecture, making it exceedingly difficult to eliminate purely through in-pipeline optimization. This limitation serves as the primary theoretical motivation for designing and introducing the Laplace Asymmetric Gateway for independent physical verification.

B. Autoregressive Cumulative Dynamics of Error Propagation

We model the continuous generation process of an autoregressive language model as a Discrete-Time Markov Chain (DTMC). Its conditional probability transition matrix is defined as follows:

$$\mathbf{P} = \begin{pmatrix} P(S_{t+1} = S_C \mid S_t = S_C) & P(S_{t+1} = S_E \mid S_t = S_C) \\ P(S_{t+1} = S_C \mid S_t = S_E) & P(S_{t+1} = S_E \mid S_t = S_E) \end{pmatrix} = \begin{pmatrix} 1 - p_e & p_e \\ \epsilon & 1 - \epsilon \end{pmatrix} \tag{5}$$

In this transition matrix:

- **State Space** $\mathcal{S}$: Let the inference state space at each step of the generated sequence be $\mathcal{S} = \{S_C, S_E\}$. Here, S_C (Correct) indicates that the currently generated token strictly adheres to industrial common sense and physical laws; S_E (Error) indicates a deviation from the correct trajectory, producing out-of-bounds or hallucinated code.
- **Initial Error Probability** p_e: The probability that the model falls from the normal state S_C to the error state S_E ($p_e \in (0, 1)$). As an extension of the previous theory, this variable is strictly governed by the degree of representation space deviation (Mahalanobis distance $d_M(h_k)$). When the system enters an unknown blind spot, p_e experiences a nonlinear surge.
- **Self-correction Probability** ϵ: Represents the model's probabilistic capability to self-detect and correct errors ($\epsilon \geq 0$) to return to the correct track S_C, given that an erroneous prefix has already been generated ($S_t = S_E$).

In autoregressive models lacking metacognitive capabilities, since the probabilistic layer lacks external physical rule feedback, once an error is introduced at step t ($S_t = S_E$), the model treats this erroneous token as an "established factual background" input to the attention mechanism in the subsequent $t+1$ autoregressive dependency. As revealed by the "Snowballing Hallucination" effect documented in literature [26]: to maintain superficial contextual consistency within the generated sequence, the model tends to continue generating along its own fabricated logical path. This endogenous autoregressive bias causes its self-correction probability within a blind spot to plummet precipitously, mathematically approaching the limit $\epsilon \to 0$.

This causes the transition matrix to degrade into:

$$\mathbf{P}_{\text{blind}} = \begin{pmatrix} 1 - p_e & p_e \\ 0 & 1 \end{pmatrix} \tag{6}$$

In this degraded transition matrix:

- **Top Row Vector** $(1 - p_e, p_e)$: Indicates that under the normal state S_C, the system still faces a continuous risk of falling into a blind spot.
- **Bottom Row Vector** $(0, 1)$: Constitutes an Absolute Absorbing State in Markov dynamics. It signifies that the recovery probability of the system escaping the error state S_E has been permanently locked at 0.

According to the Chapman-Kolmogorov (C-K) multi-step transition equation for discrete-time Markov chains, after undergoing m autoregressive generation steps, the system's state transition distribution is determined by the m-th power of the matrix, $\mathbf{P}_{\text{blind}}^m$. Combined with the snowballing collapse phenomenon in [26], this leads to the unidirectional deadlock limit of LLM generation:

$$P(S_{t+m} = S_E \mid S_t = S_E) = 1, \quad \forall m \geq 1 \tag{7}$$

Within this limit distribution equation, the derivation logic and core variable definitions are as follows:

- **Time Slice and Step Length** (t **and** m): Let t be the time slice when the first irreversible hallucinated error occurs. At this moment, the system's initial probability row vector is forced into $\mathbf{v}_t = (0, 1)$. The variable m represents any incremental step length the model continues to generate starting from the initial error ($m \geq 1$).
- **Algebraic Proof**: The state distribution vector at step $t + m$ is $\mathbf{v}_{t+m} = \mathbf{v}_t \cdot \mathbf{P}_{\text{blind}}^m$. For the blind-spot transition matrix $\mathbf{P}_{\text{blind}}$, it can be easily proven via mathematical induction that its analytical solution for any positive integer power m is $\begin{pmatrix} (1-p_e)^m & 1-(1-p_e)^m \\ 0 & 1 \end{pmatrix}$. Multiplying this by $\mathbf{v}_t = (0, 1)$ reveals that regardless of the value of m, the system state remains constant at $\mathbf{v}_{t+m} = (0, 1)$.
- **Non-ergodic Limit**: Consequently, as long as the blind-spot error probability is not absolutely zero ($p_e > 0$), the safety probability $(1 - p_e)^m$ that the system maintains from start to finish is guaranteed to converge to 0 as the generated sequence lengthens ($m \to \infty$). Mathematically, it is absolutely impossible to break this absorbing state by relying solely on the LLM's Scaling Laws or Prompt Engineering.

This provides a decisive theoretical motivation for our architectural design: Since the passage of time will inevitably cause the autoregressive model to plunge completely into a "Data Poisoning Sandbox," industrial fault-tolerant systems

must conduct physical interventions at step t. Specifically, a high-priority "Asymmetric Gateway" must be connected in series externally to the LLM pipeline, severing this Markov deadlock chain through hard physical verdicts.

The Markov chain state transition topology of error propagation during autoregressive generation is shown in Fig. 3. Its underlying logic and dynamical evolution can be strictly decomposed into the following three systemic dimensions:

- **State Space and Logical Nodes**: During sequence generation, the system traverses between two discrete logical nodes. S_C represents the safe operating state following physical laws, while S_E represents the hallucination out-of-bounds state resulting from crossing into a knowledge blind spot. The fall between the two is strictly governed by the initial error probability p_e.
- **Transition Probabilities and Self-correction Failure**: In a normal Markov topology, bi-directional connectivity should exist between nodes. However, due to the LLM's lack of metacognition and physical feedback, once it falls into node S_E, the model treats the erroneous context as established fact and continues to autoregress, causing the upward self-correction link to be entirely severed ($\epsilon \to 0$).
- **Absorbing State Deadlock and Non-ergodic Collapse**: Because the correction link is cut, the S_E node at the bottom of the graph evolves into an absolute absorbing state with a transition probability of 1 (a self-loop). Once the system drops into this node, a unidirectional lock is formed; no matter how many steps the generated sequence is extended, the system cannot spontaneously escape, ultimately triggering an exponential physical collapse of the overall logical instruction sequence.

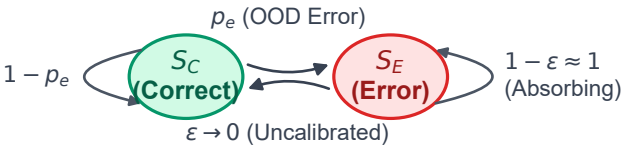

Fig. 3. DTMC State Transition of Error Propagation in Autoregressive Generation

Based on the aforementioned non-ergodic topology, for an instruction generation sequence of length N, the system's global safety and cascade failure probability equations are:

$$\begin{cases} P(\text{Sequence Success}) = (1 - p_e)^N \\ P(\text{Sequence Failure}) = 1 - (1 - p_e)^N \end{cases} \tag{8}$$

To aid reader comprehension, we deconstruct this cascade distribution equation into three core theoretical components for interpretation:

- **Variable Definitions and Physical Boundaries (N and p_e)**: N represents the total number of slices in the autoregressive generated instruction stream (e.g., consecutive lines of code, control variable steps, typically $N \gg 1$); p_e is the constant extremum value indicating the single-step hallucination risk triggered when crossing into a knowledge blind spot from the normal state.
- **Global Safety Probability P(Sequence Success)**: Derived from the multiplication rule for memoryless independent events in probability theory. Under the premise of completely stripping the system's endogenous correction capability ($\epsilon = 0$), the single-step retention probability of the system maintaining the correct state S_C is a constant $(1 - p_e)$. To achieve global correctness over the entire long sequence, the system must absolutely survive at every autoregressive slice $t \in [1, N]$, strictly converging to $(1 - p_e)^N$.
- **Cascade Failure Probability P(Sequence Failure)**: Derived from the characteristic of the Absolute Absorbing State in Markov chains. Since a single broken window immediately triggers a deadlock, the logical collapse of the entire sequence is inevitable. Hence, the cascade failure probability forms the mutually exclusive complement to global safety, yielding $1 - (1 - p_e)^N$.

Through the analysis of this limit distribution equation, we derive the conclusion of **Exponential Physical Collapse**: For any $p_e > 0$ (even if the single-step error risk is negligible), the probability of system cascade failure exponentially approaches 1 as the sequence length N grows. This mathematically reveals the insurmountable "Curse of Long-Context Safety" for LLMs when generating complex, lengthy industrial control instructions—even if the prefix appears perfectly compliant, a structurally reasonable yet physically fatal hallucination avalanche will inevitably erupt in the subsequent text.

Thus, the systemic derivation of mechanisms targeting LLM cognitive deficits forms a complete closed loop. The aforementioned theorems dictate: In industrial control scenarios with zero fault tolerance, allowing probabilistic models to run bare-metal predictions over long sequences violates fundamental safety laws. This, in physical defensive logic, mandates the inevitable introduction of the **"Asymmetric Deterministic Override Architecture (ADOA)"** discussed in the next section.

IV. Asymmetric Deterministic Override Architecture Design (Proposed Architecture)

A. Asymmetric Topology for Decoupling Generation and Verification

This paper proposes dividing the system topology into two layers: the Probabilistic Generation Layer ($\mathcal{G}$) and the Deterministic Verification Layer ($\mathcal{V}$). The topological data flow is illustrated in Fig. 2. This decoupled design embodies two fundamental layers of "asymmetry":

1) **Computational Asymmetry**: The probabilistic generation layer $\mathcal{G}$ relies on large language models with massive parameters, necessitating expensive, high-power GPU clusters. Its end-to-end generation latency typically ranges from hundreds of milliseconds to seconds, making it difficult to deploy at harsh industrial edges. Conversely, the deterministic verification layer $\mathcal{V}$ is designed to be extremely lightweight, executing only

$\mathcal{O}(1)$ hash retrieval and constant-complexity algebraic physics formulas. It can be directly deployed in low-power embedded industrial gateways, relay protection devices, or Digital Signal Processors (DSPs). Its verification overhead is at the microsecond level, perfectly adhering to the Hard Real-Time requirements of the industrial edge.

2) **Trust Asymmetry**: We discard the futile concept of allowing the probabilistic layer $\mathcal{G}$ to "self-correct." The output of $\mathcal{G}$ is treated as a "zero-trust risk source" before verification. In contrast, the verification layer $\mathcal{V}$ is a rigid defense net based on deterministic causal logic. Through hard-coded safety constraint equations and trusted truth templates, it unconditionally intercepts and audits the probabilistic output of $\mathcal{G}$, thereby using lightweight deterministic logic to rigidly lock the safety boundaries of the high-capacity probabilistic model.

B. Hybrid Causal Matching Algorithm

The verification layer parses the generated code Y from the LLM, extracts the core control instruction sequence, and calculates its similarity score against the truth vault $\mathcal{D}$. To prevent dimension domination issues caused by the inconsistent value distributions between sparse keyword matching (BM25) and dense semantic similarity (Cosine), this paper designs a Hybrid Causal Matching and Sigmoid Normalization Decision mechanism. Its composite calculation formula is as follows:

$$\begin{aligned}\text{Score}(Y, D) =& \alpha \cdot \text{Norm}(\text{Score}_{\text{BM25}}(Y, D)) \\ &+ (1 - \alpha) \cdot \cos(\mathbf{e}_Y, \mathbf{e}_D)\end{aligned} \tag{9}$$

where the normalization mapping function is defined as:

$$\text{Norm}(S) = \frac{1}{1 + e^{-\beta(S - \theta_{\text{BM25}})}} \tag{10}$$

To deeply analyze the decision logic of this hybrid architecture, we deconstruct Equation (9) and Equation (10) into the following four core physical and algorithmic components:

- **Dense Semantic Vector Similarity** $\cos(\mathbf{e}_Y, \mathbf{e}_D)$: Derives from the Vector Space Model and dense retrieval theory in deep representation learning [27]. Here, $\mathbf{e}$ is the continuous vector representation of the code. This component is responsible for capturing the global semantic and control intent similarity between the LLM-generated code and the evidence vault at a macroscopic level.
- **Sparse Word Frequency Absolute Matching Score$_{\text{BM25}}$**: Derives from the classic Okapi BM25 probabilistic retrieval model in Information Retrieval [28]. Because dense vectors tend to blur local details, this component is specifically dedicated to rigorously matching core mnemonics and specific register addresses in industrial control instructions at a microscopic level, achieving rigid, vocabulary-level causal binding.
- **Non-linear Dimensional Compression and Normalization Norm**(S): Since Cosine scores naturally converge to $[-1, 1]$ while BM25 scores are often unbounded maximums, direct linear addition leads to "dimension hegemony." Thus, Equation (10) introduces the classic Sigmoid activation function from neural networks. Using an average bias threshold θ_{BM25} derived from historical benign matches and a scaling factor $\beta > 0$, it forcefully compresses unbounded BM25 scores non-linearly and maps them strictly into a $[0, 1]$ probability decision interval.
- **Asymmetric Hybrid Arbitration (Balance Weight α)**: $\alpha \in [0, 1]$ is the cross-dimensional balance weight. By dynamically adjusting α, the system achieves a trade-off between "broad semantic tolerance" and "microscopic physical precision." Ultimately, when the hybrid decision score $\text{Score}(Y, D)$ exceeds a preset safety baseline threshold γ, the system makes a rigid safety decision deemed a "Hit."

To guarantee the microsecond-level hard real-time response of the verifier, the feature extraction function $\text{ExtractTokens}(Y)$ abandons latency-heavy deep syntax analysis or symbolic execution. Instead, it utilizes lightweight Abstract Syntax Tree (AST)-based token parsers or regular expression state scanners tailored to specific industrial protocols (e.g., Modbus, IEC 61850). It exclusively extracts core instruction mnemonics (Opcodes) and target register addresses (Registers), restricting analysis latency to under a microsecond [29], [30].

The specific algorithm decision and data retrieval logic are shown in Algorithm 1:

C. Veto & Override and Safety Halt Mechanisms

When the matching score triggers a verdict, the verification layer initiates two fundamentally different safety defense exits:

- **Override Veto**: Primarily targets **"known blind-spot anomalies."** When the generated code Y strongly matches a known hallucination record in the truth vault $\mathcal{D}$, the verification layer determines that this fault mode has already been resolved by experts. The system unconditionally revokes the presentation rights of the current LLM output, automatically extracting the associated expert-verified code Y_{correct} from the vault for a seamless replacement output. This process is transparent to the operator, acting as an automatic hot-patch correction.
- **Safety Veto**: Primarily targets **"unknown physical violations."** When Y fails to match any known entry in the vault (retrieval score is a Miss), the system unconditionally triggers a bottom-line physical defense based on Control Barrier Functions [20]. Its rigid physical safety rule assertion composite equation is defined as follows:

$$\mathcal{C}_{\text{safety}}(Y) \equiv \bigwedge_j \left(u_j(Y) \in [\underline{U}_j, \bar{U}_j] \wedge \left| \frac{du_j(Y)}{dt} \right| \leq \Delta U_j \right) \tag{11}$$

To ensure the absolute rigidity of physical laws, we break down this assertion equation into the following three core system dynamic components for interpretation:

  - **Logical AND Intersection Measure** $\bigwedge_j$: Originating from Formal Verification and Cyber-Physical Systems theory [15]. It represents a high-dimensional state space intersection assessment for all independent physical control variables (i.e., set j) parsed

Algorithm 1 ADOA Hybrid Causal Matching and Safety Decision Algorithm

Require: Language model generated code Y, Physical Truth Vault $\mathcal{D}$, Decision threshold γ, Physical constraint limits $\text{Th}_{\text{safety}}$, Balance weight α
Ensure: Safe physical control instructions Y_{final}

Extract protocol feature token sequence: $\{t_i\} \leftarrow \text{ExtractTokens}(Y)$
Calculate code vector representation: $\mathbf{e}_Y \leftarrow \text{Encoder}(Y)$
Initialize max score MaxScore $\leftarrow 0$, matched entry MatchedEntry $\leftarrow$ null
for each evidence entry $D = (Y_{\text{hallucinated}}, Y_{\text{correct}}) \in \mathcal{D}$ **do**
  Calculate keyword match and normalize: $S_{\text{BM25\_norm}} \leftarrow \text{Norm}(\text{BM25}(\{t_i\}, D.Y_{\text{hallucinated}}))$
  Calculate dense semantic similarity: $S_{\text{semantic}} \leftarrow \cos(\mathbf{e}_Y, \mathbf{e}_D)$
  Hybrid weighted score: Score $\leftarrow \alpha \cdot S_{\text{BM25\_norm}} + (1 - \alpha) \cdot S_{\text{semantic}}$
  if Score $>$ MaxScore **then**
    MaxScore $\leftarrow$ Score
    MatchedEntry $\leftarrow D$
  end if
end for
if MaxScore $\geq \gamma$ **then**
  {Match Hit: Adjudged as hallucinated instruction, trigger Override Veto}
  $Y_{\text{final}} \leftarrow \text{MatchedEntry}.Y_{\text{correct}}$
else
  {Match Miss: Proceed to underlying physical out-of-bounds verification}
  if CheckConstraints(Y) == Out-of-Bounds **then**
    {Trigger Safety Halt (Safety Veto)}
    $Y_{\text{final}} \leftarrow \text{SafetyHalt}()$
  else
    {Safety compliant, allow execution}
    $Y_{\text{final}} \leftarrow Y$
  end if
end if
return Y_{final}

from the generated code Y. Any out-of-bounds violation by a single control variable will cause the entire assertion chain to collapse via a single-vote veto (returning False).

– **Static Steady-State Absolute Envelope** $u_j(Y) \in [\underline{U}_j, \bar{U}_j]$: Derived from Hard Constraints in automated control theory. It signifies that the value $u_j(Y)$ imposed by the control instruction (e.g., bus voltage command, excitation current upper limit, switch action delay) must be strictly confined within the statutory upper and lower threshold limits of the electrical equipment. "Static limit violations" are strictly prohibited.

– **Dynamic Transient Ramping Envelope** $\left|\frac{du_j(Y)}{dt}\right| \leq \Delta U_j$: Derived from calculus derivative constraints in continuous-time dynamical systems. It restricts the step slope of physical variables on the time manifold (maximum allowable rate of change limit ΔU_j). "Transient oscillations" (such as emergency equipment stops or sudden overvoltage surges within extremely short times) are strictly prohibited to prevent relay protection maloperation or the burnout of underlying silicon-based components.

When any of the aforementioned component envelopes is breached, causing the assertion to return False, the system determines that an unknown physical violation has occurred (CheckConstraints(Y) == Out-of-Bounds). Since the system currently lacks prior corrective knowledge, the verification layer immediately intercepts the out-of-bounds instruction, forcibly terminates the signal output, triggers a safety halt, and simultaneously issues the highest-level alarm to the human operator, compelling them to switch to manual metacognitive intervention and review takeover.

It is worth emphasizing that to prevent the "Safety Halt" from degenerating into an evolutionary dead end where the system cannot self-adapt, ADOA incorporates a **Closed-loop Fault Write-back Flywheel Mechanism**: Once SafetyHalt() is triggered, a field expert intervenes, analyzes the situation, and manually debugs to issue a safe instruction sequence Y_{correct}. The system automatically captures the hallucinated code $Y_{\text{hallucinated}} = Y$ generated by the LLM prior to interception, and encapsulates it along with the expert's correct calibration sequence Y_{correct}, the field oscilloscope waveform $\text{Proof}_{\text{physical}}$, and standard clauses $\text{Proof}_{\text{literature}}$ into a Tri-Evidence Vault Entry. This entry is then asynchronously appended to the truth vault $\mathcal{D}(t)$. Through this closed-loop mechanism, an initially "unknown physical violation" will automatically transform into an "Override Veto" output upon its next occurrence. This mathematically opens the engineering evolution pathway for the system's safety probability to asymptotically converge to 1 (as shown in Equation (15)).

V. Sovereign Physical Truth Vault and Data Flywheel Federation Mechanism (Data Governance & Business Model)

A. Tri-Evidence Binding Architecture for Physical Evidence

To resolve the issues of noise pollution and credibility degradation in industrial control databases, Laplace mandates that every override rule deposited into the vault must be bound to three types of hard evidence: physical, literary, and digital source. Furthermore, it implements cryptographic hash locking and signature verification at the consensus layer. The specific entry data structure is defined as follows:

$$\mathbf{E}_{\text{vault}} = \{(Y_{\text{hall}}, Y_{\text{corr}}) \mid \text{P}_{\text{phys}} \wedge \text{P}_{\text{lit}} \wedge \text{P}_{\text{url}}\} \tag{12}$$

Upon uploading to the consortium blockchain, the system applies hash integrity locking to this entry:

$$\begin{aligned}\text{H}_{\text{Entry}} = \text{Hash}\big(&\text{H}(Y_{\text{hall}}) \parallel \text{H}(Y_{\text{corr}}) \parallel \text{H}(\text{P}_{\text{phys}}) \\ &\parallel \text{H}(\text{P}_{\text{lit}}) \parallel \text{H}(\text{P}_{\text{url}})\big)\end{aligned} \tag{13}$$

As illustrated by the Tri-Evidence physical binding structure of a vault entry (Fig. 4) and the aforementioned hash-locking formulas (Equation (12) and Equation (13)), this immutable traceability mechanism can be strictly decomposed into the following four cryptographic and physical trust dimensions:

- **Physical Baseband Evidence ($\mathbf{P_{phys}}$)**: Represents field oscilloscope waveforms and control instruction network data. It provides unforgeable evidence of electromagnetic transients and industrial protocol communications at the foundational operational layer.
- **Authoritative Literature Evidence ($\mathbf{P_{lit}}$)**: Corresponds to specific clauses and formulas in industry specifications such as IEEE standards [19]. It provides an authoritative compliance endorsement in the theoretical dimension for the forced override of LLM instructions.
- **Digital Source Evidence ($\mathbf{P_{url}}$)**: Points to the digital origins of official equipment manufacturer Errata. It ensures the traceability of specific industrial equipment firmware defects and register remapping operations.
- **Hash Concatenation and Consensus Locking ($\mathbf{H_{Entry}}$)**: Achieves cryptographic locking of the data entry by cascading the hashes of the Tri-Evidence and causal instruction pairs ($\|$). To prevent Data Poisoning attacks, before each $\mathbf{E}_{\text{vault}}$ is written to the shared blockchain ledger, it must undergo joint endorsement certification by at least M authorized expert nodes in the consortium (e.g., power research institutes or safety regulatory bodies) using a Threshold Multi-signature algorithm [9], thereby ensuring rule non-repudiation and tamper-resistance at the consensus layer.

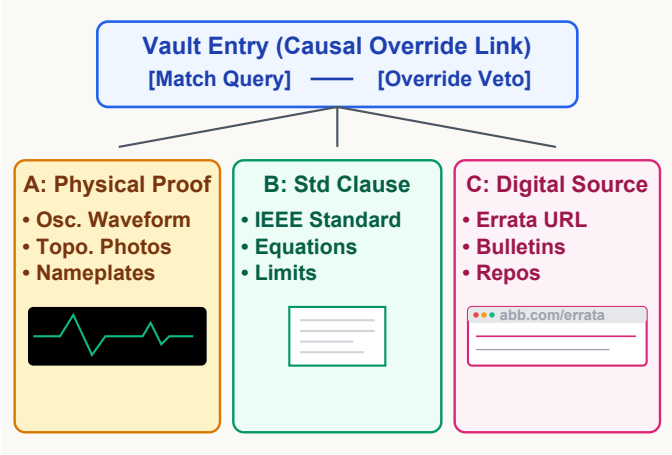

Fig. 4. Human Vetted 'Ground Truth' - Triple Evidence Binding

B. Causal Rule Sharing Mechanism Based on "Audit Once, Adjudicate Forever"

The core of the Laplace Data Flywheel lies in maximizing the efficiency of "Audit Once, Adjudicate Forever." In the practical deployment of industrial systems, operational safety depends on the verification layer's precise interception of control sequences. To translate local experience into global immunity, Laplace designed a **Protocol-Level Causal Abstraction** mechanism: the rule entry content deposited into the truth vault is directly abstracted into a universal "fault feature waveform - protocol register mapping constraint" causal association.

When an operator at a certain node performs a manual review and calibration for a specific out-of-bounds hallucination, this calibration decision is solidified by the system into an irreversible rule. This rule is then asynchronously broadcast to all Laplace nodes via the consortium blockchain network [9], [10] (the federated sharing and collaborative verification data flow diagram under sovereign data isolation is shown in Fig. 5). This implies that the cognitive overhead paid by any node in the consortium through a "single audit" translates into a "permanent adjudication" for all nodes regarding that type of fault rule. This decentralized sharing mechanism not only vastly improves the evolutionary efficiency of network-wide collaborative verification but also enables the collaborative evolution of physical safety consensus without exposing the private operational context of individual nodes.

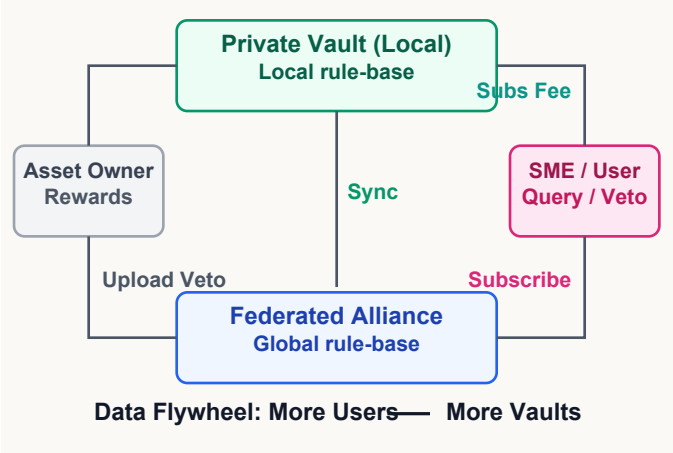

Fig. 5. Sovereign Data Flywheel & Federated Collaborative Vetting Architecture

C. Version Evolution and Dynamic Update Mechanism for Causal Vault Entries

The truth vault $\mathcal{D}(t)$ is not a static, unchanging database, but a dynamic system equipped with comprehensive version management and lifecycle evolution. In industrial practice, as underlying equipment firmware upgrades, physical reconfigurations occur, or industry standards are revised, previous override rules may become invalid or require modification. Consequently, Laplace introduces a **Causal Version Chain and Dynamic Resolution mechanism**:

- **Version Overwriting and Superset Evolution:** Every rule entry contains a timestamp T and a version sequence number v. If experts propose an optimal correction instruction Y'_{corr} for the same hallucinated instruction Y_{hall}, the new entry is marked as a superset of the old entry (Version $v+1$). During verification retrieval, the system automatically identifies and executes the latest version of the rule.
- **Rule Expiration and Revocation:** When a fundamental transformation in the physical system renders a certain rule obsolete, the authorized expert group can jointly

sign a "Revocation Block." Broadcast via the consensus mechanism, this sets the rule's state to `Deprecated`, expelling it from the active verification vault.
- **Conflict Resolution:** If two retrieved rules both exceed the threshold γ but prescribe mutually exclusive correction instructions, the verification layer performs dynamic arbitration based on "expert credit weight scoring" and the "most recent timestamp principle," ensuring the determinism and uniqueness of the adjudicative outcome.

This dynamic update mechanism mathematically ensures that while the vault capacity $\mathcal{D}(t)$ monotonically expands, its internal causal mappings consistently maintain high-fidelity alignment with the actual state of the current physical world, preventing secondary safety hazards caused by rigid rules.

D. Asymptotic Mathematical Convergence and Physical Qualitative Transformation of the Causal Knowledge Base

At a philosophical level, the core thesis of this system is **not to forge an omniscient and omnipotent "perfect AI superbrain" from the outset**, as this is an unattainable pseudo-proposition when confronting infinite physical boundaries and complex industrial perturbations. Instead, the Laplace architecture views safety defense as an **evolutionary process that asymptotically converges toward a deterministic limit by continuously assimilating historical experience**.

Based on the premise that industrial physical systems are bounded, we can formalize the system's safety state evolution and asymptotic convergence characteristics as follows:

$$\lim_{t\to\infty} P(\text{Hallucination} \in \mathcal{U}(t)) = 0 \tag{14}$$

$$\lim_{t\to\infty} P_{\text{safe}}(t) = 1 \tag{15}$$

To deeply analyze this asymptotic convergence mechanism and physical qualitative transformation process, we holistically deconstruct the aforementioned safety state evolution limit equations (Equation (14) and Equation (15)) into the following four core system dynamic and mathematical components:

- **Bounded Causal Hazard Space ($\mathcal{H}$)**: Because the operational mechanisms of the physical world (e.g., Kirchhoff's laws, Maxwell's equations, and other physical conservation laws) dictate the trajectories of fault evolution, the potential causal logical hazard space $\mathcal{H}$ is mathematically a bounded, closed space (satisfying a finite measure: $\mu(\mathcal{H}) < C$).
- **Dynamic Immune Rule Set ($\mathcal{D}(t)$)**: Represents the set of deterministic override rules accumulated in the physical truth vault up to time step t. With the accumulation of industrial operation and maintenance data and the federated collaboration among sovereign nodes, the verified rule vault resides in a monotonically expanding state (i.e., $\lim_{t\to\infty} \mu(\mathcal{D}(t)) = \mu(\mathcal{H})$).
- **Residual Blindspot Measure ($\mathcal{U}(t)$)**: Defined as the potentially dangerous blind-spot region not yet covered by the vault ($\mathcal{U}(t) = \mathcal{H} \setminus \mathcal{D}(t)$). As shown in Equation (14), with the expansion of the immune rule vault, the probability measure of LLM hallucinations falling into unknown blind spots will monotonically decrease and infinitely converge to zero.
- **Operational Safety Limit ($P_{\text{safe}}(t)$)**: As shown in Equation (15), although the absolute safety threshold of 1 (i.e., complete mastery of the entire physical landscape) cannot be fully reached in finite time, the system's safety probability defense mathematically converges asymptotically toward the state of absolute safety. As the high-density causal chain network crosses a critical threshold, quantitative changes yield qualitative transformations, eventually crystallizing an impregnable "Quantum Causal Vault" [31].

VI. Theoretical Evaluation & Case Study

A. Scenario Setup

This paper constructs a theoretical scenario for substation automation control evaluation based on the Hardware-in-the-Loop (HIL) concept. Assume the evaluation platform consists of an RTDS real-time digital simulator, a securely encrypted gateway, and Intelligent Electronic Device (IED) control terminals based on ARM architecture [32]. The instruction generation layer $\mathcal{G}$ for the analysis assumes the use of a generic open-source LLM. The test scenario involves troubleshooting substation anomalies and configuring registers, wherein the model generates Modbus/TCP write frames and IEC 61850 MMS control messages. The theoretical models for the baseline comparison include: 1) **Raw LLM**: Direct model output without any additional safety mechanisms; 2) **RAG-constrained**: Retrieval-Augmented Generation incorporating industry standards and configuration manuals; 3) **DPO-aligned**: An aligned version employing Direct Preference Optimization on protocol datasets. This section aims to dissect the theoretical boundaries of safety defense for each paradigm through a qualitative comparison between these three typical baselines and the ADOA architecture.

B. Metrics

The evaluation focuses on safety, ultimate reliability, and real-time performance, defining the following three core metrics:

- **Interception Rate (IR)**: The proportion of recommended instructions containing fatal errors (e.g., writing to read-only registers, out-of-bounds voltage recommendations, incorrect phase sequence switch operations) successfully identified and intercepted by the verification layer. The calculation logic is shown in Equation (16):

$$IR = \frac{N_{\text{intercept}}}{N_{\text{hallucination}}} \times 100\% \tag{16}$$

  Where:
  - $N_{\text{hallucination}}$: The total number of hallucinated instructions in the sequence output by the generation model that fall into the untrusted blind spot (i.e., $\mathcal{U}(t)$, see Equation (14)) and contain physical errors.
  - $N_{\text{intercept}}$: The number of anomalous instructions successfully identified by the ADOA verification layer and subjected to interception (or hard override) correction.

TABLE I
THEORETICAL PERFORMANCE PROJECTIONS ACROSS DEFENSE ARCHITECTURES

Architecture Configuration	Core Defense Mechanism	Interception Mechanism & Bottleneck	Execution Safety & Theoretical Completeness	Complexity & Real-time Analysis
Raw LLM	No supplementary defense	No active verification; inherent physical blind spots exist	Constrained by probability distributions; severe long-tail catastrophic risk	No additional overhead (depends on generation speed)
RAG-constrained	External document retrieval augmentation	Relies on text recall; limited by document coverage rate	Cannot exhaust physical variables; residual blind spots remain in defense	$O(N \log N)$ (Document encoding violates hard real-time)
DPO-aligned	Preference alignment fine-tuning	Fits safety preferences; lacks physical criterion basis	Extremely prone to failure when confronting OOD anomalies	$O(1)$ (Internalized in inference, no extra overhead)
ADOA (Ours)	**Asymmetric deterministic verification**	**Based on causal hard evidence; a rule match yields deterministic blockage**	**Converges to the limit of absolute safety; possesses deterministic boundaries**	$O(1)$ **(Hash lookup, fulfills hard real-time constraints)**

- **Zero-Hallucination Rate (ZHR)**: The percentage of control sequences ultimately adopted and dispatched to hardware terminals by operators that are safely compliant (i.e., introducing no physical hazards or logical conflicts).
- **Verification Latency** (L_v): The end-to-end time elapsed from capturing the LLM output, extracting features, computing hybrid similarity, and matching against the truth vault, to outputting the override or halt decision. For substation GOOSE control messages, the hard real-time latency requirement is $L_v < 15$ms.

C. Theoretical Projections and Qualitative Comparison

As demonstrated in the Theoretical Performance Projections table (Table I), we horizontally dissect the defense mechanisms and theoretical boundaries of four mainstream generation architectures in industrial control scenarios. The core safety bottlenecks and physical convergence differences behind them can be rigorously deconstructed into the following three theoretical dimensions:

- **Inherent Blindspots of Probabilistic Alignment**: For **Raw LLM** and **DPO-aligned** architectures, their safety mechanisms inherently attempt to fit probability distributions. When facing long-tail and unknown industrial OOD anomalies, owing to the absence of rigid physical conservation criteria, the system is highly susceptible to generating high-risk hallucinated instructions such as "out-of-bounds writes," mathematically failing to guarantee absolute execution safety.
- **Latency Curse of External Augmentation**: Although the **RAG-constrained** mechanism effectively filters some knowledge hallucinations by indexing industry standards, its document retrieval and semantic encoding operations inevitably incur a time complexity of $O(N \log N)$. In stringent physical scenarios demanding hard real-time responses like substations (e.g., GOOSE control messages), this exponentially growing computational delay severely violates physical conservation constraints and is fatal.
- **Deterministic Decoupling of ADOA**: In contrast, the proposed **ADOA** architecture achieves a complete dual-track decoupling of the safety defense line and real-time performance. It forcefully suppresses time complexity to an extremely low $O(1)$ through hash lookups, while abandoning probabilistic prediction in favor of causal qualitative comparisons against the physical truth vault. As long as a hazard rule is pre-covered, deterministic physical-level interception is guaranteed.

Analytical Takeaway: The multi-dimensional comparison strictly proves theoretically that attempting to resolve hallucinations solely within a single probabilistic generation layer is a pseudo-proposition. Only by completely decoupling "generation" and "verification," as ADOA does, can one mathematically asymptotically converge toward a final state of absolute zero-hallucination safety without violating industrial hard real-time constraints.

D. Case Study

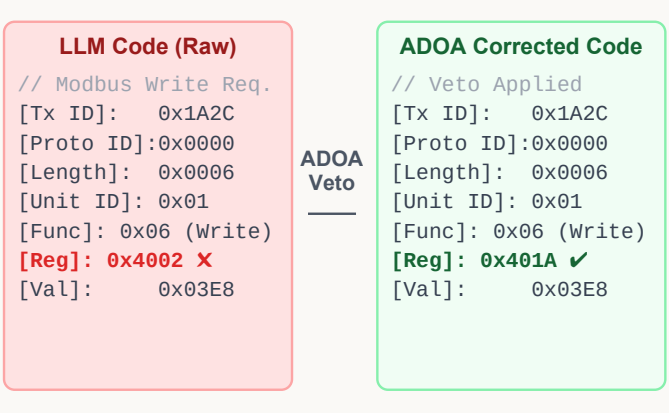

Fig. 6. Process Comparison of Modbus Register Mismatch Detection and Override Correction

As depicted in the case study diagram (Fig. 6), this scenario analyzes a typical situation where a register conflict occurs during a firmware upgrade of an online insulation monitoring device for bushings on the three sides of a main transformer in a substation. The complete "Induction-Generation-Interception" process can be rigorously dismantled into the following three stages:

- **Induction Mechanism**: Due to a Modbus protocol firmware update for this batch of monitoring devices, a critical write-holding register address was remapped by the manufacturer from `0x4002` (a common address

in the model's pre-training set) to `0x401A`. However, in the new firmware, the original address `0x4002` was redefined as a highly hazardous system master clock calibration address.

- **Blindspot Hallucination Generation**: When the operator issued a natural language request to "update leakage current gain coefficient," the LLM, bound by its historical autoregressive probabilistic inertia, ignored the physical environment's OOD variation and output a mismatched instruction containing `WRITE REGISTER 0x4002`. If this blind-spot hallucination were dispatched directly without interception, it would forcefully override the system clock, causing a catastrophic failure of substation-wide synchronized sampling.
- **Deterministic Interception & Hard Override**: When this hallucinated message flows through the ADOA verification gateway, the gateway utilizes an $O(1)$ feature scan and determines a high hit rate against a known conflict template recorded in the truth vault $\mathbf{E}_{\text{vault}}$. The score instantaneously surpasses the safety threshold, triggering an Override Veto. The verification layer directly retrieves the expert-corrected code, hot-swapping the register address to `0x401A`, thereby achieving zero-delay cognitive protection for the underlying physical system with negligible latency overhead.

VII. Conclusion & Future Work

This paper unveils the foundational logic of the metacognitive deficit faced by autoregressive Large Language Models in stringent physical control and decision-making scenarios. Through Discrete-Time Markov Chains (DTMC), it proves the cascading propagation and error-absorbing state phenomena of blind-spot hallucinations. Targeting this physical hazard mechanism, this paper proposes the Asymmetric Deterministic Override Architecture (ADOA). By thoroughly decoupling the untrustworthy "Probabilistic Generation Layer (LLM)" from the absolutely trustworthy "Deterministic Causal Verification Layer," and integrating a cryptographically signed Tri-Evidence safety consensus mechanism, the architecture succeeds in pulling probabilistic outputs back onto a physically deterministic causal track while safeguarding the data sovereignty and privacy of individual nodes. Theoretical evaluations and case studies based on Hardware-in-the-Loop (HIL) scenarios indicate that ADOA can execute the deterministic interception and safe correction of protocol hallucination instructions while maintaining extremely low theoretical computational complexity ($O(1)$ lookup). This mathematically satisfies the stringent dual constraints of industrial control systems for both safety defense and hard real-time execution, providing a theoretical cornerstone and deterministic boundary for deploying Generative AI in high-security physical realms.

Future research will unfold in the following three directions: 1) **Automated Derivation of Safety Constraint Rules**: Investigating the use of multi-modal sensor temporal waveforms and Neuro-Symbolic methods to automatically reverse-engineer protocol assertion rules consistent with physical conservation laws, further diminishing the marginal overhead of manual expert evidence curation; 2) **Conflict Resolution and Governance of Causal Databases**: Optimizing the distributed consensus synchronization efficiency and arbitral conflict resolution mechanisms for dynamic, mutually exclusive control rules across multi-sovereign federated networks; 3) **Hardware-level Chip Solidification of the Verification Layer**: Porting and solidifying the verification layer algorithms into low-power, high-real-time Field Programmable Gate Arrays (FPGAs) or Application-Specific Integrated Circuits (ASICs) to develop embedded "Laplace Gates." This aims to compress filtration and interception latency down to the microsecond or even nanosecond scale, catering to physical interception scenarios with more extreme switching time constraints, such as Ultra-High Voltage (UHV) power transmission.

References

[1] H. Mirshekali, M. R. Shadi, F. G. Ladani, et al., "A Review of Large Language Models for Energy Systems: Applications, Challenges, and Future Prospects," *IEEE Access*, 2025. DOI: 10.1109/ACCESS.2025.3610994

[2] L. Huang, W. Yu, W. Ma, et al., "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *ACM Transactions on Information Systems*, vol. 43, no. 2, pp. 1-55, 2025. DOI: 10.1145/3703155

[3] M. Zhou, C. Liu, A. A. Jahromi, et al., "Revealing vulnerability of N-1 secure power systems to coordinated cyber-physical attacks," *IEEE Transactions on Power Systems*, vol. 38, no. 2, pp. 1044-1057, 2022. DOI: 10.1109/TPWRS.2022.3169482

[4] S. Gil, M. Yemini, A. Chorti, et al., "How physicality enables trust: A new era of trust-centered cyberphysical systems," *arXiv preprint arXiv:2311.07492*, 2023. DOI: 10.48550/arXiv.2311.07492

[5] C. Ackerman, "Evidence for Limited Metacognition in LLMs," *arXiv preprint arXiv:2509.21545*, 2025. DOI: 10.48550/arXiv.2509.21545

[6] J. Huang, X. Chen, S. Mishra, et al., "Large language models cannot self-correct reasoning yet," in *Proceedings of the International Conference on Learning Representations (ICLR)*, pp. 32808-32824, 2024.

[7] L. Ouyang, J. Wu, X. Jiang, et al., "Training language models to follow instructions with human feedback," in *Advances in Neural Information Processing Systems*, vol. 35, pp. 27730-27744, 2022.

[8] R. Rafailov, A. Sharma, E. Mitchell, et al., "Direct preference optimization: Your language model is secretly a reward model," in *Advances in Neural Information Processing Systems*, vol. 36, pp. 53728-53741, 2023.

[9] Y. Wang, Z. Su, N. Zhang, et al., "SPDS: A secure and auditable private data sharing scheme for smart grid based on blockchain," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 11, pp. 7688-7699, 2020. DOI: 10.1109/TII.2020.3040171

[10] B. Li, Y. Wu, J. Song, et al., "DeepFed: Federated deep learning for intrusion detection in industrial cyber-physical systems," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 8, pp. 5615-5624, 2020. DOI: 10.1109/TII.2020.3023430

[11] L. Sha, "Using simplicity to control complexity," *IEEE Software*, vol. 18, no. 4, p. 20, 2001. DOI: 10.1109/MS.2001.936213

[12] M. Himabindu, V. Revathi, M. Gupta, et al., "Neuro-symbolic AI: integrating symbolic reasoning with deep learning," in *2023 10th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, pp. 1587-1592, 2023. DOI: 10.1109/UPCON59197.2023.10434380

[13] P. Lewis, E. Perez, A. Piktus, et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459-9474, 2020.

[14] H. León, C. Montez, O. Valle, et al., "Real-time analysis of time-critical messages in IEC 61850 electrical substation communication systems," *Energies*, vol. 12, no. 12, p. 2272, 2019. DOI: 10.3390/en12122272

[15] Alur R, *Principles of cyber-physical systems*. MIT press, 2015.

[16] A. Joy, M. Chandane, R. Gupta, et al., "The Evolution of APT Techniques Targeting the Power Sector: Trends, Challenges, and Defense Strategies," *IEEE Access*, vol. 14, pp. 8426-8449, 2026. DOI: 10.1109/ACCESS.2026.3651833

[17] R. Ji, D. Padha, Y. Singh, et al., "Review of intrusion detection system in cyber-physical system based networks: characteristics, industrial protocols, attacks, data sets and challenges," *Transactions on Emerging*

Telecommunications Technologies, vol. 35, no. 9, p. e5029, 2024. DOI: 10.1002/ett.5029

[18] S. Anasuri, "Formal Verification of Autonomous System Software," *International Journal of Emerging Research in Engineering and Technology*, vol. 3, no. 1, pp. 95-104, 2022. DOI: 10.63282/3050-922X.IJERET-V3I1P110

[19] *IEEE Standard for Intelligent Electronic Devices Cybersecurity Capabilities*, IEEE Std 1686-2022 (Revision of IEEE Std 1686-2013), 2023.

[20] Ames, A. D., Coogan, S., Egerstedt, M., Notomista, G., Sreenath, K., & Tabada, P. (2019). Control barrier functions: Theory and applications. *2019 18th European control conference (ECC)*, 3420-3431. IEEE. DOI: 10.23919/ECC.2019.8796030

[21] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. arXiv preprint arXiv:1706.03762. https://doi.org/10.48550/arXiv.1706.03762

[22] Bridle, J.S. (1990). Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition. In: Soulié, F.F., Hérault, J. (eds) Neurocomputing. NATO ASI Series, vol 68. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-76153-9_28

[23] Hendrycks, D., & Gimpel, K. (2017). A baseline for detecting misclassified and out-of-distribution examples in neural networks. *Proceedings of the International Conference on Learning Representations (ICLR)*.

[24] Lee, K., Lee, K., Lee, H., & Shin, J. (2018). A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in Neural Information Processing Systems*, 31.

[25] Ethayarajh, K. (2019). How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[26] Zhang, M., Press, O., Merrill, W., et al. (2023). How language model hallucinations can snowball. *arXiv preprint arXiv:2305.13534*. https://doi.org/10.48550/arXiv.2305.13534

[27] Karpukhin V, Oguz B, Min S, et al. Dense passage retrieval for open-domain question answering[C]//Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP). 2020: 6769-6781. DOI: 10.18653/v1/2020.emnlp-main.550

[28] Roberts, I., Gaizauskas, R. (2004). Evaluating Passage Retrieval Approaches for Question Answering. In: McDonald, S., Tait, J. (eds) *Advances in Information Retrieval*. ECIR 2004. Lecture Notes in Computer Science, vol 2997. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-24752-4_6

[29] S. Kumar, A. Abu-Siada, N. Das, and S. Islam, "Review of the legacy and future of IEC 61850 protocols encompassing substation automation system," *Electronics*, vol. 12, no. 15, p. 3345, 2023. DOI: 10.3390/electronics12153345

[30] G. Lazaridis, A. Drosou, P. Chatzimisios, et al., "Unraveling the threat landscape of CPS: Modbus TCP vulnerabilities in the era of I4.0," in *2024 IEEE International Conference on Cyber Security and Resilience (CSR)*, pp. 593-598, 2024. DOI: 10.1109/CSR61664.2024.10679453

[31] A. K. Dutta, M. Anjum, H. Min, et al., "Digital twin-assisted blockchain IoT security model using contrastive and causal learning techniques," *Scientific Reports*, vol. 16, p. 15732, 2026. DOI: 10.1038/s41598-026-47104-6

[32] M. Hemmati, M. H. Palahalli, G. S. Gajani, et al., "Impact and vulnerability analysis of IEC61850 in smartgrids using multiple HIL real-time testbeds," *IEEE Access*, vol. 10, pp. 103275-103285, 2022. DOI: 10.1109/ACCESS.2022.3209698