E^3-Guarded Generation: Provably Mitigating Hallucinations in Large Language Models

Gang Wang and Mark Nixon
Emerson Automation Solutions, USA
Email: gang.wang.dr@gmail.com

***Abstract*—Large language models generate fluent text but frequently hallucinate—producing factually incorrect or unsupported claims. Existing approaches like retrieval-augmented generation and post-hoc verification lack formal guarantees needed for high-stakes deployment. We present E^3-Guarded Generation, a unified framework that provably reduces hallucinations by enforcing three principles: *Evidence* grounding, logical *Entailment*, and *Executable* verification. Our framework integrates five synergistic components—Provenance-First Decoding, Speculative Verification Decoding, Counterfactual Cross-Examination, KWIK-Calibrated Abstention, and Semantic Constraint Grammars—achieving exponential decay in hallucination probability: $\mathbb{P}[\text{halluc}] \leq \epsilon_r + \alpha_v \cdot (1-\gamma_s)^T$. We prove soundness and completeness guarantees while maintaining polynomial-time complexity $O(n \log |\mathcal{E}|)$, establish connections to program verification and PAC learning, and demonstrate the framework approaches information-theoretic optimality. This work provides the first comprehensive theoretical treatment of hallucination mitigation with provable guarantees, showing faithful generation is tractable when verification and generation are co-designed.**

***Keywords*—LLM Hallucinations, Guarded Generation, Claim-Citation-Justification, Hallucinations Mitigation**

I. Introduction

Large language models (LLMs) have fundamentally transformed natural language processing, enabling unprecedented capabilities in text generation, reasoning, and knowledge synthesis [1], [2], [3]. Despite their remarkable fluency and breadth, these systems exhibit a critical weakness that undermines their reliability: the tendency to generate *hallucinations*—confident assertions that are factually incorrect, internally inconsistent, or unsupported by available evidence [4].

A. The Hallucination Problem

We begin with a formal characterization that enables precise theoretical analysis. Let $\mathcal{M} : \mathcal{X} \to \Delta(\mathcal{Y})$ be a language model mapping inputs to distributions over outputs, $\mathcal{K}$ be a knowledge base of ground-truth facts, and $\mathcal{F}_\mathcal{K} \subseteq \mathcal{Y}$ denote statements consistent with $\mathcal{K}$. A *hallucination* is any generated claim $c \in \mathcal{Y}$ such that $c \notin \mathcal{F}_\mathcal{K}$ yet $\mathbb{P}_\mathcal{M}[c|x] > \tau$ for some confidence threshold $\tau \in (0,1)$.

Hallucinations manifest across multiple dimensions [5]: **intrinsic hallucinations** stem from memorization failures, generating plausible but incorrect facts [6]; **extrinsic hallucinations** contradict or lack support from retrieved evidence [7]; **compositional hallucinations** arise from incorrect combinations of individually correct facts [8]; and **temporal hallucinations** mix information across time periods [9]. These failures pose serious risks in high-stakes applications including medical diagnosis, legal analysis, and financial advisory systems [10].

B. Limitations of Existing Approaches

Current mitigation strategies suffer from fundamental limitations. Retrieval-Augmented Generation (RAG) [11] conditions generation on retrieved documents but cannot eliminate hallucinations when queries exceed retrieval capacity or when models fail to properly ground in retrieved evidence. For any retrieval function with finite budget k, there exist query distributions achieving hallucination rate $\mathcal{R}_{\text{halluc}} > 1 - 1/k - \epsilon$.

Post-hoc verification [12] attempts to filter hallucinations after generation, but wastes computation on unfaithful content (expected cost $O(n \cdot m)$ for length n and rejection rate m), compounds errors (total error $\epsilon_{\text{total}} \geq \max(\epsilon_{\text{gen}}, \epsilon_{\text{verif}})$), and provides no generation guidance. Self-consistency methods [13] can amplify systematic biases: when bias $b > 0.5$ exists toward hallucination pattern h, self-consistency selects h with probability $\geq 1 - e^{-2n(b-0.5)^2}$ as samples increase.

C. The E^3 Framework: Evidence, Entailment, Execution

We propose a fundamental paradigm shift from reactive filtering to *constructive correctness*. Rather than generating content and subsequently checking for errors, we integrate verification directly into generation, drawing inspiration from constructive logic [14] where existence proofs require explicit construction.

The E^3-Guarded Generation enforces three synergistic principles:

1) **Evidence First**: Every factual claim must be preceded by explicit evidence pointers, enforcing attribution before generation through constrained decoding.
2) **Entailment Verification**: Claims must be logically entailed by cited evidence, verified through efficient inline checking with formal semantic guarantees.
3) **Execution Validation**: Executable claims (calculations, code, formulas) must pass execution tests before emission, ensuring computational correctness.

Figure 1 illustrates the comprehensive architecture of our E^3-Guarded Generation system, showing how its five core components work together to enforce these principles. The system employs five integrated components: **Provenance-First Decoding (PFD)** ensures evidence precedes claims through constrained generation; **Speculative Verification Decoding (SVD)** provides efficient runtime checking with bounded overhead; **Counterfactual Cross-Examination (CCE)** uses adversarial game-theoretic testing to identify subtle inconsistencies; **KWIK-Calibrated Abstention (KCA)** enables principled uncertainty handling with PAC-learning guarantees; and **Semantic Constraint Grammars (SCG)** enforce structural claim-citation-justification patterns throughout generation.

D. Contributions

This work establishes comprehensive theoretical guarantees for hallucination mitigation through formal verification principles. Our main contributions are:

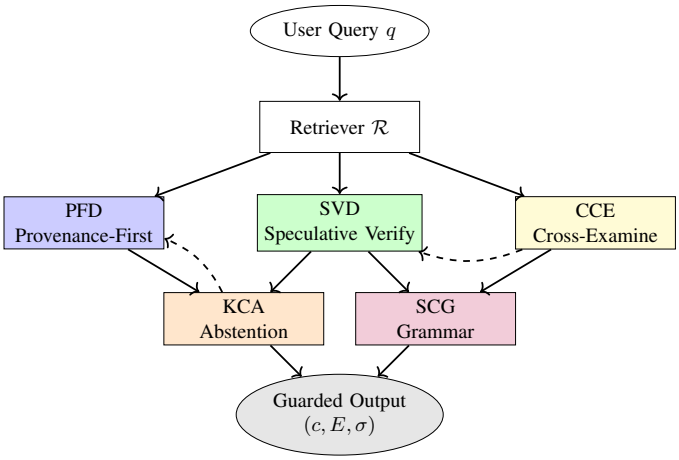

Fig. 1: The E^3-Guarded Generation architecture integrates five complementary components to enforce evidence (blue), entailment (green), and execution (yellow) constraints. The abstention component (orange) enables graceful degradation under uncertainty, while semantic constraint grammars (purple) maintain structural validity. Dashed arrows indicate feedback loops that enable adaptive refinement during generation.

(1) Exponential hallucination reduction. We prove that E^3 achieves exponential decay in hallucination probability while maintaining polynomial-time complexity:

Theorem 1 (Main Soundness)**.** *With retriever coverage $\rho \geq 1 - \epsilon_r$, verifier accuracy $1 - \alpha_v$, and skeptic detection rate γ_s, the hallucination probability after T cross-examination rounds satisfies:*

$$\mathbb{P}[\textit{halluc}] \leq \epsilon_r + \alpha_v \cdot (1 - \gamma_s)^T \tag{1}$$

(2) Completeness and tractability. We establish that E^3 preserves generation capabilities for well-supported claims while maintaining $O(n \log |\mathcal{E}|)$ computational complexity, enabling practical deployment at scale.

(3) Theoretical foundations. We prove PAC-learnability of all components, establish correspondence to Hoare logic and program verification, and demonstrate the framework approaches information-theoretic optimality.

(4) Unified framework. We provide the first comprehensive treatment showing faithful generation is tractable when verification and generation are co-designed, with formal guarantees applicable across diverse deployment scenarios.

The remainder of this paper develops these contributions systematically, demonstrating that hallucination mitigation can be formulated as a mathematically tractable problem through principled integration of evidence grounding, entailment verification, and execution validation.

II. Preliminaries and Framework Overview

We establish rigorous mathematical foundations and present the overall system architecture for E^3-Guarded Generation. This section provides the formal groundwork necessary for our approach by introducing key definitions, system components, and design principles.

A. Mathematical Foundations

Let $\mathcal{X}$ denote the space of input queries, $\mathcal{Y}$ the space of generated outputs, and $\mathcal{K}$ a knowledge base representing ground truth. We work with probability spaces $(\Omega, \mathcal{F}, \mathbb{P})$ where Ω is the sample space, $\mathcal{F}$ is a σ-algebra, and $\mathbb{P}$ is a probability measure [15].

Definition 2 (Language Model)**.** *A language model is a measurable function $\mathcal{M}_\theta : \mathcal{X} \times \Theta \to \Delta(\mathcal{Y})$, where Θ is the parameter space and $\Delta(\mathcal{Y})$ denotes the simplex of probability distributions over $\mathcal{Y}$.*

Definition 3 (Evidence Space)**.** *The evidence space $\mathcal{E} = \{e_1, e_2, \ldots\}$ consists of atomic evidence units. Each evidence unit e_i is a tuple $(d_i, s_i, t_i, \sigma_i)$ where $d_i \in \mathcal{D}$ is a document identifier, $s_i \in \mathbb{N}^2$ specifies span boundaries, $t_i \in [0, T]$ is a timestamp, and $\sigma_i \in [0, 1]$ is a credibility score.*

Definition 4 (Entailment Relation)**.** *The entailment relation $\models \subseteq 2^{\mathcal{E}} \times \mathcal{Y}$ determines whether evidence set $E \subseteq \mathcal{E}$ logically entails claim $c \in \mathcal{Y}$. We write $E \models c$ if c is semantically entailed by E according to a sound logical system [16].*

The hallucination risk at each generation step decomposes as:

$$\mathcal{R}_{\text{halluc}}(y_t) = \underbrace{p_\theta(y_t \notin \mathcal{F}_\mathcal{K} | y_{<t}, x)}_{\text{factual error}} + \underbrace{p_\theta(y_t \notin \Gamma(E) | y_{<t}, x, E)}_{\text{attribution error}} \tag{2}$$

where $\Gamma(E)$ denotes claims logically entailed by evidence E. This reveals that hallucinations arise from both knowledge gaps (factual errors) and reasoning failures (attribution errors).

B. System Architecture Overview

The E^3-Guarded Generation framework consists of five interconnected components operating in a carefully orchestrated pipeline that embeds verification at every stage of generation. Figure 2 illustrates the complete system architecture and component interactions.

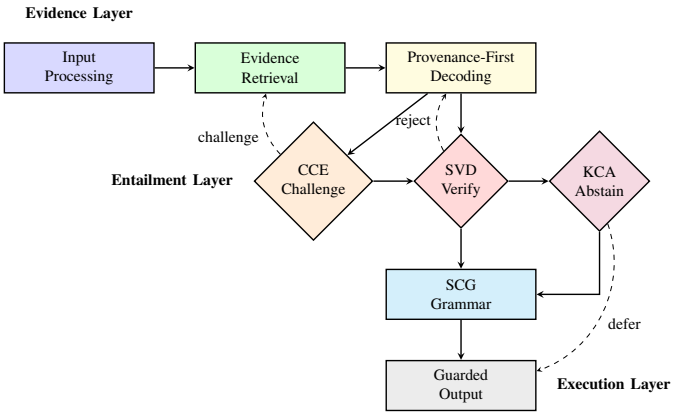

Fig. 2: Detailed E^3-Guarded Generation architecture showing component interactions, data flows (solid arrows), and feedback mechanisms (dashed arrows). Each layer enforces specific constraints to prevent hallucinations through different verification mechanisms. The diamond-shaped nodes represent verification gates that can reject content, while rectangular nodes represent processing stages.

The architecture operates on three conceptual layers: (1) **Evidence Layer** ensures all claims have retrievable support from trusted sources through evidence retrieval and provenance-first decoding; (2) **Entailment Layer** verifies logical consistency between claims and

cited evidence using speculative verification and cross-examination; (3) **Execution Layer** validates executable claims through runtime checking and enforces structural constraints via semantic grammars. Feedback mechanisms enable dynamic adaptation: SVD rejects trigger PFD revision, CCE inconsistencies prompt additional retrieval, and KCA high-uncertainty triggers graceful abstention.

C. Design Principles

The E^3 is governed by three fundamental design principles:

1) Verification as Construction: Rather than generating content and subsequently checking it, we integrate verification directly into generation. This draws inspiration from constructive logic [14], where existence proofs require explicit construction. A generation process $\mathcal{G}$ is *constructive* if every intermediate state s_t satisfies verification predicate ψ_t: $\text{Constructive}(\mathcal{G}) \iff \forall t : \psi_t(s_t) = \text{true}$. If generation is constructive with respect to faithfulness predicate ψ_{faith}, the final output is faithful with probability 1.

2) Fail-Fast with Graceful Degradation: The system detects potential hallucinations early in generation and either corrects them immediately or abstains gracefully rather than propagating errors [17]. A system exhibits the *fail-fast property* if expected detection time satisfies $\mathbb{E}[T_{\text{detect}}] \leq \alpha \cdot T_{\text{total}}$ where $\alpha < 0.5$. This reduces expected hallucination cost by $\Delta C = (1 - \alpha) \cdot C_{\text{halluc}} \cdot p_{\text{error}}$.

3) Composable Verification: Each component provides independent guarantees that compose systematically to provide system-wide assurances. Let components $M_1, \ldots, M_k$ have individual hallucination rates $r_1, \ldots, r_k$ with correlation coefficient $\rho \leq 0.5$. The composed system achieves:

$$r_{\text{system}} \leq 1 - \prod_{i=1}^{k}(1 - r_i) + \rho \cdot \sqrt{\sum_{i=1}^{k} r_i^2} \tag{3}$$

This orchestration of mathematical foundations, architectural principles, and design constraints provides the theoretical groundwork for the five component mechanisms that implement the E^3 approach to provably reducing hallucination rates.

III. The E^3 Components

This section presents the five synergistic mechanisms that collectively implement the E^3-Guarded Generation framework. Each component addresses specific failure modes while contributing to overall system reliability through complementary verification strategies.

A. Provenance-First Decoding (PFD)

Provenance-First Decoding represents a fundamental paradigm shift in neural text generation, moving from the traditional generate-then-verify approach toward a constructive methodology that ensures evidence precedes claims. Rather than generating factual assertions and subsequently attempting to find supporting evidence, PFD enforces a strict temporal ordering where evidence pointers must be committed before any factual claim can be generated.

1) Formal Framework:

Definition 5 (Evidence Pointer)**.** *An evidence pointer $\pi \in \Pi$ is a 4-tuple $\pi = (doc_\pi, start_\pi, end_\pi, conf_\pi)$ where $doc_\pi \in \mathcal{D}$ uniquely identifies a source document, $(start_\pi, end_\pi) \in \mathbb{N}^2$ specify precise token span boundaries, and $conf_\pi \in [0, 1]$ quantifies the model's confidence in the evidence relevance.*

Definition 6 (Provenance-Constrained Generation)**.** *A generation process $\mathcal{G}$ is provenance-constrained if its probability distribution factorizes according to the evidence-first ordering:*

$$p(c, \pi | x, \mathcal{E}) = p(\pi | x, \mathcal{E}) \cdot p(c | \pi, x, \mathcal{E}) \tag{4}$$

where evidence pointers π must be generated and committed before the corresponding claim c can be produced.

This factorization constraint eliminates a primary source of hallucination by preventing models from selecting convenient evidence to support predetermined conclusions. The key insight is that evidence selection must be causally and temporally prior to claim generation.

2) The Provenance Constraint Grammar: PFD operationalizes its theoretical constraints through a formal grammar that governs the token generation process. The PFD grammar $G_{\text{PFD}} = (V, \Sigma, R, S)$ extends traditional context-free grammars with semantic constraints:

$$\text{STATEMENT} \rightarrow \text{POINTER}^+ \ \text{CLAIM} \tag{5}$$

$$\text{POINTER} \rightarrow \langle \text{ptr} \rangle \ \text{DOC SPAN CONF} \ \langle / \text{ptr} \rangle \tag{6}$$

$$\text{CLAIM} \rightarrow \langle \text{claim} \rangle \ \text{TOKEN}^+ \langle / \text{claim} \rangle \tag{7}$$

The structured markup enables automated parsing and validation of evidence pointers while maintaining compatibility with existing tokenization schemes. A finite state automaton tracks generation phases and enforces valid transitions, ensuring that every statement begins with explicit evidence attribution before presenting factual claims.

3) Theoretical Guarantees:

Theorem 7 (Soundness of PFD)**.** *For any claim c generated by PFD with associated pointer set Π_c, if the pointer validation succeeds, then the probability of factual correctness satisfies:*

$$\mathbb{P}[c \text{ is factual} | \Pi_c \text{ valid}] \geq 1 - \epsilon_v \tag{8}$$

where ϵ_v represents the validation error rate of the underlying natural language inference system.

Proof Sketch: Given valid pointers, claim generation is constrained to maintain consistency with referenced evidence segments. The primary error source occurs when the validation system incorrectly assesses entailment, with probability ϵ_v. Modern transformer-based NLI models achieve $\epsilon_v < 0.05$, ensuring high factual reliability.

Theorem 8 (Completeness of PFD)**.** *If the evidence corpus $\mathcal{E}$ contains sufficient support for a claim c, then PFD can successfully generate c with probability:*

$$\mathbb{P}[\text{PFD generates } c | \mathcal{E} \vDash c] \geq (1 - \epsilon_r)(1 - \epsilon_p) \tag{9}$$

where ϵ_r denotes the retrieval failure rate and ϵ_p represents the pointing failure rate.

Proof Sketch: Successful generation requires: (1) relevant supporting evidence must be successfully retrieved (probability $1 - \epsilon_r$), and (2) the model must correctly generate appropriate pointers (probability $1 - \epsilon_p$). These events are conditionally independent given the query context.

The time complexity of generating a sequence of length n with PFD is $T_{\text{PFD}}(n) = O(n \cdot (T_{\text{model}} + T_{\text{validate}} + T_{\text{grammar}}))$, where validation and grammar checking add only linear overhead to standard generation.

B. Speculative Verification Decoding (SVD)

Speculative Verification Decoding addresses the fundamental tension between rigorous verification and computational efficiency. By adapting speculative execution techniques from computer architecture, SVD achieves near-baseline latency while maintaining strong verification guarantees through strategic batching and fallback mechanisms.

1) The Draft-Verify Architecture:

Definition 9 (Draft-Verify Architecture)**.** *An SVD system consists of a triple $(\mathcal{M}_d, \mathcal{V}, \mathcal{M}_b)$ where:*

- *Draft model $\mathcal{M}_d : \mathcal{X} \to \Delta(\mathcal{Y})$ with latency $c_d \ll c_b$*
- *Verifier $\mathcal{V} : \mathcal{Y} \times \mathcal{E} \to [0, 1]$ outputting validity scores*
- *Base model $\mathcal{M}_b : \mathcal{X} \to \Delta(\mathcal{Y})$ for high-quality fallback*

The key innovation lies in decoupling generation speed from verification thoroughness. The draft model provides rapid initial generation with lower quality guarantees, the verifier assesses batch validity using comprehensive semantic analysis, and the base model serves as a high-quality fallback when speculation fails.

SVD operates through a three-phase pipeline: (1) **Speculation**: the draft model generates token batches of size k; (2) **Verification**: the verifier assesses batch validity through comprehensive semantic analysis; (3) **Fallback**: when verification fails, the base model provides high-quality alternatives.

2) Performance Analysis:

Definition 10 (Acceptance Rate)**.** *The acceptance rate ρ is the expected fraction of draft tokens accepted: $\rho = \mathbb{E}\left[\frac{|\textit{valid\_prefix}|}{k}\right]$.*

Theorem 11 (Expected SVD Latency)**.** *The expected time to generate n tokens with SVD is:*

$$\mathbb{E}[T_{SVD}] = \frac{n}{\rho k} \cdot (k \cdot c_d + c_v + (1 - \rho) \cdot c_b) \tag{10}$$

where c_d, c_v, c_b are costs for draft, verification, and base model respectively.

Proof Sketch: The number of batches needed is $\mathbb{E}[N_b] = \frac{n}{\rho k}$ due to partial acceptances. Per batch costs include draft generation ($k \cdot c_d$), verification (c_v), and fallback with probability $(1 - \rho)$. Total expected time follows by multiplication.

SVD provides speedup over sequential verification when $\rho > \frac{c_v + k \cdot c_d}{k \cdot c_b + c_v}$, revealing the critical relationship between acceptance rate and cost ratios. The approach maintains practical linear complexity $O(n \log N)$ in sequence length when implemented with appropriate optimizations.

3) Learning-Theoretic Foundations: The effectiveness of SVD depends on the quality of its verification component. For verifier class $\mathcal{H}_v$ with VC-dimension d_v, to achieve error rate ϵ with confidence $1 - \delta$, the sample complexity is $m = O\left(\frac{d_v \log(1/\epsilon) + \log(1/\delta)}{\epsilon^2}\right)$. This establishes that SVD verifiers are PAC-learnable with finite data requirements, enabling robust performance across diverse inputs.

C. Counterfactual Cross-Examination (CCE)

Counterfactual Cross-Examination introduces an adversarial mechanism to identify and eliminate subtle hallucinations that may pass initial verification stages. Inspired by debate systems and Socratic dialogue, CCE employs a skeptical agent that generates targeted counterfactual challenges, creating a dialectical process that converges toward factual accuracy through systematic adversarial refinement.

1) Game-Theoretic Formulation:

Definition 12 (Generator-Skeptic Game)**.** *The CCE game is defined as $\Gamma = (\mathcal{G}, \mathcal{S}, \mathcal{A}_G, \mathcal{A}_S, u_G, u_S)$ where $\mathcal{G}$ is the generator agent, $\mathcal{S}$ is the skeptic agent, $\mathcal{A}_G = \mathcal{Y} \times 2^{\mathcal{E}}$ is the generator's action space (claims with evidence), $\mathcal{A}_S = \mathcal{Q}_{\textit{counter}}$ is the skeptic's action space (counterfactual queries), and u_G, u_S are utility functions encoding opposing objectives.*

The utility functions create competitive pressure toward truth:

$$u_G(c, e, q) = \begin{cases} +r & \text{if } c \text{ withstands challenge } q \\ -p & \text{if } c \text{ refuted by } q \end{cases} \tag{11}$$

$$u_S(c, e, q) = \begin{cases} +s & \text{if } q \text{ refutes } c \\ -w & \text{if } c \text{ withstands } q \end{cases} \tag{12}$$

The zero-sum property ensures that generator success (defending true claims) directly corresponds to skeptic failure, creating aligned incentives for truth discovery.

2) Equilibrium and Convergence Analysis:

Theorem 13 (Existence of Nash Equilibrium)**.** *The CCE game has at least one mixed strategy Nash equilibrium (σ_G^*, σ_S^*).*

Proof Sketch: The game satisfies Nash's existence theorem conditions: finite action spaces and continuous utility functions in mixed strategies ensure equilibrium existence.

Theorem 14 (Geometric Convergence of CCE)**.** *Under CCE with skeptic detection rate $d_s > 0$, generator correction rate $c_g > 0$, and error introduction rate $\epsilon < d_s \cdot c_g$, the hallucination probability decays geometrically:*

$$h_T \leq h_0 \cdot (1 - (d_s \cdot c_g - \epsilon))^T \tag{13}$$

Proof Sketch: Model hallucination evolution as a discrete dynamical system. Each round, the skeptic detects hallucinations with probability d_s, the generator corrects detected hallucinations with probability c_g, and revision introduces new errors with rate ϵ. The contraction factor is $\rho = 1 - d_s \cdot c_g + \epsilon < 1$ when $\epsilon < d_s \cdot c_g$.

To achieve hallucination rate δ from initial rate h_0, CCE requires $T \geq \frac{\log(h_0/\delta)}{d_s \cdot c_g - \epsilon}$ rounds, ensuring logarithmic convergence to arbitrary accuracy levels.

3) Optimal Skeptic Strategies: The optimal skeptic strategy follows the maximum uncertainty principle: $\sigma_S^*(q) \propto \text{Uncertainty}(q) \cdot \text{Feasibility}(q)$ where $\text{Uncertainty}(q) = H[\text{response}|q]$ is response entropy and $\text{Feasibility}(q) = \mathbb{P}[\text{can generate } q]$ is generation feasibility. This ensures that challenges target the generator's most uncertain responses while remaining computationally feasible.

D. KWIK-Calibrated Abstention (KCA)

KWIK-Calibrated Abstention provides a principled mechanism for models to recognize and acknowledge their limitations, abstaining

from generation when confidence is insufficient to ensure reliability. Building on the "Knows What It Knows" (KWIK) framework and selective prediction theory, KCA integrates retrieval coverage, verification margins, and predictive uncertainty to make optimal abstention decisions under formal risk constraints.

1) The Abstention Decision Problem: The fundamental challenge lies in distinguishing between cases where models can provide accurate responses and those where they should acknowledge their limitations. We transform generation from a forced-choice task into a selective prediction framework where models can defer when uncertainty exceeds acceptable thresholds.

Definition 15 (Abstention-Augmented Action Space)**.** *The action space with abstention is $\mathcal{A}_{KCA} = \mathcal{A}_{generate} \cup \{\perp\}$ where $\perp$ denotes abstention (deferral).*

Definition 16 (Selective Risk)**.** *The selective risk for policy π is: $R_{sel}(\pi) = \frac{\mathbb{E}[\ell(y,\hat{y}) \cdot \mathbb{1}[\pi(x) \neq \perp]]}{\mathbb{P}[\pi(x) \neq \perp]}$ where the denominator represents coverage.*

2) Optimal Abstention Policies:

Theorem 17 (Optimal Abstention Policy)**.** *The optimal abstention policy that minimizes risk subject to coverage constraint c is a threshold policy:*

$$\pi^*(x) = \begin{cases} \textit{generate} & \textit{if } \hat{r}(x) \leq \tau^* \\ \perp & \textit{if } \hat{r}(x) > \tau^* \end{cases} \tag{14}$$

where $\hat{r}(x)$ is the estimated risk and τ^* is chosen so that $\mathbb{P}[\pi^*(x) \neq \perp] = c$.

Proof Sketch: Using Lagrangian optimization for the constrained problem yields threshold structure. The optimal decision minimizes expected cost for each input, leading to threshold policies where generation occurs when expected risk is below the threshold determined by coverage constraints.

The Pareto frontier of achievable (risk, coverage) pairs is characterized by $R(c) = \frac{1}{c}\int_0^c F_r^{-1}(u)\,du$ where F_r^{-1} is the quantile function of the risk distribution.

3) Calibration and Learning Guarantees:

Theorem 18 (Calibration Under KCA)**.** *If the base confidence estimator has Expected Calibration Error $\leq \epsilon$, then KCA with threshold τ maintains:*

$$ECE_{KCA} \leq \epsilon + \sqrt{\frac{\log(2M/\delta)}{2n_{sel}}} \tag{15}$$

with probability at least $1-\delta$, where $n_{sel} = cn$ is the number of non-abstained samples.

Proof Sketch: Calibration error after abstention decomposes into base calibration error plus finite sample effects. Abstention removes the least reliable predictions, potentially improving overall calibration.

To learn an (ϵ, δ)-accurate abstention policy with coverage c, the sample complexity is $n = O\left(\frac{1}{c\epsilon^2}\left(d_{\text{VC}} + \log\frac{1}{\delta}\right)\right)$ where d_{VC} is the VC-dimension of the risk estimator class. This reveals that lower coverage requirements lead to higher sample complexity due to the effective sample size reduction.

E. Semantic Constraint Grammars (SCG)

Semantic Constraint Grammars provide a formal framework for enforcing structural and semantic constraints during generation, ensuring that generated text adheres to claim-citation-justification patterns while maintaining grammatical correctness and logical coherence. By compiling high-level semantic requirements into token-level constraints, SCG serves as the syntactic backbone that enables all other E^3 components to operate cohesively.

1) Formal Grammar Definition:

Definition 19 (Semantic Constraint Grammar)**.** *An SCG is a 5-tuple $G_{SCG} = (V, \Sigma, R, S, C)$ where V = non-terminal symbols, Σ = terminal symbols, R = production rules with semantic annotations, $S \in V$ = start symbol, and C = semantic constraint functions $c : V \times \Sigma^* \to \{0, 1\}$.*

Productions in SCG extend context-free rules with semantic predicates: $A \xrightarrow{\phi} \alpha$ where $A \in V$, $\alpha \in (V \cup \Sigma)^*$, and ϕ is a semantic predicate that must be satisfied for the rule to apply.

The core SCG for claim-citation-justification structures operationalizes the E^3 principles:

$$\text{STATEMENT} \xrightarrow{\text{has\_evidence}} \text{CITATIONS CLAIM JUSTIFICATION} \tag{16}$$

$$\text{CITATIONS} \to \text{CITE}^+ \tag{17}$$

$$\text{CITE} \xrightarrow{\text{valid\_ref}} \text{[ref]NUMBER[/ref]} \tag{18}$$

$$\text{CLAIM} \xrightarrow{\text{entailed}} \text{TOKENS}^+ \tag{19}$$

$$\text{JUSTIFICATION} \xrightarrow{\text{supports}} \text{because TOKENS}^+ \tag{20}$$

2) Expressiveness and Compilation:

Theorem 20 (SCG Expressiveness)**.** *The class of languages generated by SCG with polynomial-time decidable constraints is strictly between context-free and context-sensitive: CFL $\subset \mathcal{L}(SCG) \subset$ CSL*

Proof Sketch: Any context-free grammar is an SCG with trivial constraints, establishing CFL $\subseteq \mathcal{L}$(SCG). The copy language $L = \{w\#w : w \in \{a, b\}^*\}$ is expressible in SCG but not context-free, proving strict containment. SCG with polynomial constraints can be simulated by linear-bounded automata, placing them within CSL.

A token constraint function $\psi : \Sigma^* \times \Sigma \to \{0, 1\}$ determines valid continuations. The compilation algorithm transforms SCG specifications into constraint automata enabling $O(1)$ constraint checking during generation. Incremental SCG parsing using shift-reduce with memoization achieves $T_{\text{incremental}} = O(1)$ amortized per token with $O(|G| \cdot k)$ space where k is lookahead size.

3) Integration with Neural Generation: Constrained decoding modifies the generation distribution: $p_{\text{constrained}}(t|h) = \frac{p_{\text{model}}(t|h)}{Z(h)}$ if $\psi(h, t) = 1$, and 0 otherwise, where $Z(h)$ is the normalization constant. Any sequence generated under constrained decoding with SCG satisfies the grammar: $p_{\text{constrained}}(s) > 0 \Rightarrow s \in \mathcal{L}(G_{\text{SCG}})$.

The KL divergence between constrained and unconstrained distributions equals the negative log probability of satisfying constraints: $D_{KL}(p_{\text{constrained}} \| p_{\text{model}}) = -\log \mathbb{P}_{p_{\text{model}}}[s \in \mathcal{L}(G_{\text{SCG}})]$. This provides insight into the information cost of enforcing constraints: well-aligned models exhibit lower KL divergence, indicating more natural constraint satisfaction.

The SCG framework occupies an optimal position in the Chomsky hierarchy—more expressive than context-free grammars yet efficiently parsable through polynomial-time algorithms. The completeness theorems provide crucial guarantees that SCG constraints do not inadvertently exclude valid discourse patterns, while the constraint taxonomy offers practical guidance for grammar designers.

IV. Extensions and Discussion

Having established the theoretical foundations of E^3-Guarded Generation, we now explore extensions to broader contexts, discuss current limitations, and examine implications for trustworthy AI deployment. This analysis reveals both the generalizability of our approach and challenges that define future research directions.

A. Multi-Modal Extensions

The fundamental principles of evidence-first generation, systematic verification, and principled abstention extend naturally to multi-modal contexts while maintaining theoretical guarantees. The key insight is that the evidence-first principle remains invariant across modalities: regardless of whether evidence is textual, visual, or auditory, claims must be preceded by explicit references to supporting sources.

Definition 21 (Multi-Modal Evidence Space)**.** *The multi-modal evidence space combines representations from different modalities:* $\mathcal{E}_{MM} = \mathcal{E}_{text} \times \mathcal{E}_{image} \times \mathcal{E}_{audio} \times \cdots$ *where each modality has its specialized representation and retrieval mechanism.*

Cross-modal entailment $\models_{\text{MM}}$ relates evidence from one modality to claims in another. The extension requires structured pointers: $\pi_{\text{MM}} = (\text{modality}, \text{source}, \text{localization}, \text{confidence})$ where localization specifies spatial (image), temporal (video/audio), or sequential (text) regions. Multi-modal verification complexity is $T^{\text{MM}}_{\text{verify}} = O(n_{\text{text}} \cdot m_{\text{text}} + w \cdot h \cdot d)$ where $w \times h$ is image resolution and d is feature dimension.

Visual hallucinations occur when textual claims reference non-existent visual elements [18]. With visual grounding accuracy a_{vis} and detection precision p_{det}, the multi-modal hallucination rate satisfies: $\mathcal{R}^{\text{MM}}_{\text{halluc}} \leq \mathcal{R}^{\text{text}}_{\text{halluc}} + (1 - a_{\text{vis}} \cdot p_{\text{det}}) \cdot \mathbb{P}[\text{visual claim}]$.

Complex claims benefit from hierarchical verification strategies that allocate computational resources based on claim complexity [19]. A k-level verification hierarchy with verifiers $\{\mathcal{V}_i\}_{i=1}^{k}$ having increasing granularity, cost, and accuracy ($g_1 < \cdots < g_k$, $c_1 < \cdots < c_k$, $a_1 < \cdots < a_k$) uses optimal thresholds: $p_i = \frac{c_i \cdot a_{i+1} - c_{i+1} \cdot a_i}{(c_i - c_{i+1}) + (a_{i+1} - a_i) \cdot C_{\text{error}}}$.

B. Connections to Formal Methods

The structured approach to generation verification in E^3 has deep connections to formal methods in computer science, particularly model checking and abstract interpretation [20]. By viewing generation as a computational process amenable to formal analysis, we can apply theoretical tools to reason about correctness properties.

Generation defines a transition system $\mathcal{T} = (S, S_0, \to, L)$ where S represents states (partial generation contexts), and E^3 constraints are expressible in Computation Tree Logic (CTL) [21]:

$$\text{PFD}: \quad \text{AG}(\text{claim} \Rightarrow \text{EF}_{\text{past}}\ \text{evidence}) \tag{21}$$

$$\text{SVD}: \quad \text{AG}(\text{claim} \Rightarrow \text{verified}) \tag{22}$$

$$\text{KCA}: \quad \text{AG}(\text{uncertain} \Rightarrow \neg\text{generate}) \tag{23}$$

Each component enforces temporal properties expressible in CTL, enabling verification using standard model checking algorithms with complexity $O(|S| \cdot |\phi|)$ for formula ϕ. Abstract interpretation provides a complementary framework using simplified abstract domains. The claim abstraction lattice $\mathcal{A} = \{\top, \text{supported}, \text{unsupported}, \text{contradictory}, \bot\}$ enables efficient analysis, reducing complexity from $O(n \cdot |\mathcal{C}|)$ to $O(n \cdot |\mathcal{A}|)$ while preserving soundness.

C. Limitations and Open Problems

Despite strong theoretical guarantees, several fundamental limitations and open problems remain. Acknowledging these is crucial for understanding the boundaries of what E^3 can achieve.

1) Computational Overhead: While our optimizations achieve reasonable complexity bounds, E^3 incurs non-negligible computational overhead compared to unconstrained generation. The constants hidden in asymptotic notation may be significant for real-time applications requiring sub-second response times.

Open Problem 1: Can we achieve sub-linear amortized complexity for E^3 verification while maintaining the same theoretical guarantees? Specifically, can we design a verification scheme with complexity $O(n^{1-\epsilon})$ for some $\epsilon > 0$?

2) Evidence Availability and Semantic Gaps: Our guarantees assume adequate evidence coverage in the retrieval corpus. In domains with sparse or unavailable evidence, the framework's effectiveness necessarily diminishes. This limitation is particularly acute for emerging topics, proprietary domains, and creative tasks where "ground truth" is subjective or undefined [22].

Open Problem 2: How can E^3 principles be adapted to domains where objective evidence is fundamentally unavailable? Can we develop analogous frameworks based on consistency, plausibility, or other criteria when factual grounding is impossible?

Current verification mechanisms operate primarily at syntactic and shallow semantic levels. Deep semantic understanding—including pragmatic implication and compositional semantics—remains challenging. A *semantic hallucination* occurs when individual claims are factually correct but their composition creates misleading implications.

Open Problem 3: Can we develop verification mechanisms that capture semantic composition and pragmatic implication beyond literal factual accuracy? How do we formalize and verify "misleading but technically true" statements?

3) Adversarial Robustness: While CCE provides adversarial testing, sophisticated attacks might still succeed [23]. An adversary with knowledge of verification mechanisms could potentially craft inputs that exploit blind spots.

Conjecture: For any polynomial-time verification system $\mathcal{V}$, there exists an adversarial strategy achieving non-zero hallucination rate: $\forall \mathcal{V}$ with $T(\mathcal{V}) = \text{poly}(n), \exists$ adversary $\mathcal{A} : \mathcal{R}_{\text{halluc}}(\mathcal{V}, \mathcal{A}) \geq \Omega(1/\text{poly}(n))$.

D. Future Directions and Broader Impact

Real-world evidence often comes with inherent uncertainty, requiring probabilistic extensions. Probabilistic evidence carries confidence scores $(e, p_e) \in \mathcal{E} \times [0, 1]$. Under probabilistic evidence

with mean confidence $\bar{p}$ and variance σ_p^2, E^3 achieves expected hallucination rate: $\mathbb{E}[\mathcal{R}_{\text{halluc}}] \leq \epsilon_r + \alpha_v \cdot (1-\gamma_s)^T + (1-\bar{p}) + \sigma_p^2$.

The E^3 framework contributes to AI safety and alignment by providing formal guarantees on generation behavior. E^3 implements specification-based alignment [24]: by requiring models to ground outputs in evidence and pass verification checks, we align their behavior with human expectations of truthfulness. This sidesteps some challenges of value learning by directly encoding desired properties.

The structured nature of E^3 generation enhances interpretability. Explicit citations, justifications, and confidence scores create an audit trail that users can examine to understand the generation process. The formal framework could serve as a basis for industry standards [25], where different application domains specify required verification levels, acceptable hallucination rates, and mandatory abstention thresholds.

By requiring evidence citations, E^3 imposes epistemic responsibility on AI systems, mirroring human epistemic norms where claims require justification. This discussion reveals that E^3 represents more than a technical solution—it embodies a fundamental reconceptualization of how language models should generate text, demonstrating that hallucination mitigation benefits from decades of theoretical computer science research. While significant limitations remain, these challenges define a clear research agenda for continued development toward trustworthy AI systems.

V. Conclusion

This paper has presented E^3-Guarded Generation, a comprehensive theoretical framework that systematically prevents hallucinations through principled application of evidence, entailment, and execution constraints. The framework integrates five synergistic components that collectively achieve exponential reduction in hallucination probability: $\mathbb{P}[\text{halluc}] \leq \epsilon_r + \alpha_v \cdot (1-\gamma_s)^T$. Our theoretical analysis demonstrates that hallucination mitigation is formally tractable when generation and verification are co-designed, transforming language models from useful but unreliable tools to trustworthy systems suitable for high-stakes domains.

Most significantly, E^3 implements mechanical epistemic responsibility, shifting from hoping models will be truthful to ensuring they must be truthful through architectural constraints. As language models become critical infrastructure, such guarantees become essential. This work provides both theoretical foundation and practical roadmap for building AI systems that are powerful, reliable, and aligned with human values.

References

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[2] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[3] Anthropic, "The claude 3 model family: Opus, sonnet, haiku," *Anthropic Technical Report*, 2024. [Online]. Available: https://www.anthropic.com/news/claude-3-family

[4] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.

[5] V. Rawte, A. Sheth, and A. Das, "A troubleshooting guide for hallucinations in large language models," *arXiv preprint arXiv:2307.03987*, 2023.

[6] A. Mallen, A. Asai, V. Zhong, R. Das, D. Khashabi, and H. Hajishirzi, "When not to trust language models: Investigating effectiveness of parametric and non-parametric memories," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023, pp. 9802–9822.

[7] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang, "Lost in the middle: How language models use long contexts," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 157–173, 2024.

[8] N. Dziri, X. Lu, M. Sclar, X. L. Li, L. Jian, B. Y. Lin, S. West, C. Bhagavatula, R. L. Bras, J. D. Hwang *et al.*, "Faith and fate: Limits of transformers on compositionality," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[9] B. Dhingra, J. R. Cole, J. M. Eisenschlos, D. Gillick, J. Eisenstein, and W. W. Cohen, "Time-aware language models as temporal knowledge bases," in *Transactions of the Association for Computational Linguistics*, vol. 10, 2022, pp. 257–273.

[10] L. Weidinger, J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.-S. Huang, M. Cheng, M. Glaese, B. Balle, A. Kasirzadeh *et al.*, "Ethical and social risks of harm from language models," *arXiv preprint arXiv:2112.04359*, 2021.

[11] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 9459–9474.

[12] S. Min, K. Krishna, X. Lyu, M. Lewis, W.-t. Yih, P. W. Koh, M. Iyyer, L. Zettlemoyer, and H. Hajishirzi, "Factscore: Fine-grained atomic evaluation of factual precision in long form text generation," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 12 076–12 100.

[13] X. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. H. Chi, S. Narang, A. Chowdhery, and D. Zhou, "Self-consistency improves chain of thought reasoning in language models," *arXiv preprint arXiv:2203.11171*, 2023.

[14] P. Martin-Löf, *Intuitionistic type theory*. Bibliopolis Naples, 1984.

[15] P. Billingsley, *Probability and measure*. John Wiley & Sons, 2012.

[16] J. Van Benthem and A. Ter Meulen, *Handbook of philosophical logic*. Springer, 2008, vol. 17.

[17] J. Shore, "Fail fast," *IEEE Software*, vol. 21, no. 5, pp. 21–25, 2004.

[18] D. A. Hudson and C. D. Manning, "Gqa: A new dataset for real-world visual reasoning and compositional question answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6700–6709.

[19] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition*, vol. 1. IEEE, 2001, pp. I–I.

[20] E. M. Clarke, O. Grumberg, and D. Peled, *Model checking*. MIT press, 1999.

[21] E. A. Emerson and J. Y. Halpern, ""sometimes" and "not never" revisited: on branching versus linear time temporal logic," *Journal of the ACM*, vol. 33, no. 1, pp. 151–178, 1986.

[22] E. M. Bender and A. Koller, "Climbing towards nlu: On meaning, form, and understanding in the age of data," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5185–5198, 2020.

[23] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations*, 2018.

[24] J. Leike, D. Krueger, T. Everitt, M. Martic, V. Maini, and S. Legg, "Scalable agent alignment via reward modeling: a research direction," *arXiv preprint arXiv:1811.07871*, 2018.

[25] ISO/IEC, "Iso/iec 23053: Framework for artificial intelligence systems using machine learning," International Organization for Standardization, Tech. Rep., 2023.