CONSTRAINING PROBABILITY WITH LOGIC: A SPECTRUM FROM STATISTICAL ALIGNMENT TO STRUCTURAL GUARANTEE

Anonymous authors
Paper under double-blind review

ABSTRACT

Hallucination mitigation for large language models (LLMs) is approaching the point of diminishing marginal returns: prompt engineering, retrieval-augmented generation (RAG), and reinforcement learning from human feedback (RLHF) can drive error probabilities extremely low, yet they cannot break through a **structural non-zero lower bound**. We trace this limit to an **epistemological tension** between **probabilistic likelihood** (Likelihood; used here in the broad sense of "probabilistic plausibility," not the statistical likelihood function) and **logical necessity** (Necessity): the mathematical nature of standard Softmax with finite logits and without hard masking (for any token, $P(t) > 0$ holds) conflicts irreconcilably with logical validity (which requires exact exclusion of certain outputs). To address this, we propose the **Values→Support continuum** as an analytic framework. Along the **information path** (changing probability values), a mathematical ceiling is unavoidable; along the **structural path** (hard masking to reconstruct the feasible set V_{valid}), the probability of logically invalid tokens can be made exactly zero. Building on this, we construct **three layers of logical intervention**—information (empiricism), normative (rationalism), and structural (formalism)—forming a **defense in depth** from statistical enhancement to formal guarantees. Rather than providing a final engineering solution, this paper establishes a **research agenda** to shift AI reliability from **soft alignment** toward **hard guarantees**, enabling a paradigm shift in how we govern hallucinations.

1 INTRODUCTION: THE STRUCTURAL PREDICAMENT OF LLM HALLUCINATIONS

1.1 DIMINISHING RETURNS AND THE SPECTRUM FRAMEWORK

The reasoning capability of LLMs has been continually optimized through chain-of-thought prompting (CoT; Wei et al., 2022), retrieval-augmented generation (RAG; Lewis et al., 2020), and reinforcement learning from human feedback (RLHF; Ouyang et al., 2022). These improvements, however, are approaching a **boundary of diminishing marginal returns**. In particular, on the more challenging LogiConBench (Chen et al., 2026), state-of-the-art models achieve roughly **85–95%** on **discriminative tasks**, but on **enumerative tasks** the best exact accuracy is **only 34%** (under the benchmark's overall/average reporting), revealing a structural gap from "judging" to "constructing." In addition, as an independent observation and diagnostic in this paper, complex reasoning chains still frequently exhibit three critical distortions: contradiction (a conclusion conflicts with premises or intermediate propositions), reference drift (the same entity's reference slips or is replaced across the chain), and premise-free assertions (skipping necessary support and jumping statements without traceable grounding).

This predicament is not merely an engineering shortfall; it reflects a **mismatch between a mathematical mechanism and an epistemological goal**. LLMs operate in **Likelihood** (again: broad "probabilistic plausibility," not the statistical likelihood function), whereas deductive reasoning requires **Necessity**. Probabilistic language admits every non-zero possibility; logical language requires exact exclusion. Existing strategies—e.g., Logic-of-Thought (LoT; Liu et al., 2025), which reweights conditional distributions by injecting logical information into context, or Sparse Multi-Agent Debate

(SMAD; Yang et al., 2026), which improves auditability via multi-agent debate—still operate inside a probability distribution. They cannot bypass the inherent constraint of standard Softmax with **finite logits and no hard mask**: for any token, $P(t) > 0$ holds, so logically invalid outputs can only be pushed down, never excluded exactly.

We therefore reveal a **continuous spectrum from soft to hard** in logical interventions. The **information path** (changing probability values) is bounded by Softmax's mathematics and thus exhibits an **insurmountable systemic non-zero lower bound**; the **structural path** (hard masking to reconstruct the **support /** feasible set) achieves **probability exactly zero** for invalid tokens via explicit exclusion. We organize these interventions into **three layers**—information (empiricism), normative (rationalism), and structural (formalism)—as threshold regions on the spectrum, corresponding to progressive strengthening from statistical enhancement to formal guarantees.

Accordingly, this paper does not offer a final solution. Instead, it provides a diagnosis and a direction: generalizing existing *syntactic* hard constraints to *semantic/logical* constraints, and outlining a route toward the **structural endpoint**. The choice of formal systems, optimization of constraint checkers, and the realizability of open-domain constraints (which often exhibit engineering complexity close to AI-Complete) require joint work by formal logicians, neural-network engineers, and cognitive scientists. We offer this paper as an invitation to evolve "constraining probability with logic" from an engineering trick into a system architecture: not by waiting for general AI breakthroughs, but by continuously decomposing open-domain problems into decidable/verifiable restricted subdomains, and progressively expanding the coverage of the structural path.

1.2 Positioning: A Diagnostic Framework and an Epistemological Intervention

This paper is **not** an engineering optimization study, nor a compromise between symbolism and connectionism. It is an **epistemological intervention** that touches the foundational assumptions of AI.

To confront the structural malady of reasoning hallucinations, we argue that a probabilistic generation mechanism cannot natively carry logical necessity, because "highly improbable" cannot be made equivalent to "impossible." We revisit a central insight from Johan van Benthem's dynamic logic—**logic gains its force by deleting impossible states and thereby reconstructing the space of possibilities**—and translate it into a principle of **support-set pruning**. This translation requires suspending the habitual question of "how to reduce errors within a probabilistic framework," and **redefining the problem itself**: shifting from gradual statistical optimization to structural governance via **boundary setting**, making certain errors *mechanistically impossible* within formalizable fragments.

This positioning resembles a **research programme**: it does not aim to solve a single engineering problem; it changes the ontology of the problem so that certain impossibilities become unavoidable boundaries—turning the gradual optimization goal of "reducing hallucination rates" into a structural governance goal of "rebuilding verifiability on mathematical necessity." Through the Values→Support continuum, we rebuild a structural anchor for verifiability in a probability-dominant era: transforming reasoning errors from "low-probability risks" into **mechanistic exclusion zones**, and providing later technical work with a **syntax for asking questions** and a **terrain for evaluation**.

In this sense, the paper represents an epistemological reset from a **probability-optimization paradigm** to a **structural-governance paradigm**. The former assumes Softmax can ultimately bear logical reliability, so one only needs to push the probability P of correct conclusions toward 1. The latter acknowledges a mathematical boundary: hard constraints must be embedded into the generation architecture to reconstruct feasible sets. This is not a patch; it is a change in the **ontology of the problem**.

2 The Structural Tendency of Hallucinations

Mainstream views often treat LLM hallucinations as incidental noise inside an **engineering optimization space**, assuming scale, alignment, and prompting will eventually erase them. We diagnose a **structural non-zero lower bound**: not an accidental defect due to insufficient training, but a struc-

tural necessity induced by Softmax's mathematics and the epistemic presuppositions of statistical learning.

2.1 Mathematical Basis: Softmax's Incompressible Lower Bound

We start from the most basic mathematics. The core generation mechanism of LLMs relies on Softmax to convert logits into a probability distribution. Given context c, the probability of the next token t is:

$$P(t \mid c) = \frac{\exp(z_t)}{\sum_{t' \in V} \exp(z_{t'})}$$

where z_t is the logit of token t, and V denotes the full vocabulary. The **support** of this distribution is the set of values with non-zero probability mass:

$$\mathrm{supp}(P(\cdot \mid c)) := \{t \in V \mid P(t \mid c) > 0\}.$$

Under the standard Softmax setting with **finite logits and no hard mask**, $\mathrm{supp}(P(\cdot \mid c)) = V$.

This formula entails a key mathematical fact: the exponential function $\exp(x)$ maps any finite real number to a strictly positive value. No matter how negative z_t is—whether through training or prompting—$\exp(z_t) > 0$ always holds. Therefore, in standard Softmax over finite logits without masking, every token $t \in V$ has strictly positive probability:

$$\forall t \in V : \quad P(t \mid c) > 0.$$

This is not pedantry; it is a mathematical necessity. In practice, greedy decoding, temperature scaling, and soft biases such as logit bias/penalties can compress the sampling probability of invalid tokens to extremely small values (approximately 0 under finite sampling). In a single or small number of generations, this can feel indistinguishable from "impossible." Yet **"highly improbable" is not "impossible" in mathematics**.

In **long-chain reasoning**, if we approximately treat "an error at each step" as an independent event and the per-step error probability is about ε (with $\varepsilon \ll 1$), then the probability of at least one error within n steps is:

$$P_{\text{error}}(n) = 1 - (1 - \varepsilon)^n \approx n\varepsilon,$$

where the approximation holds when $\varepsilon \ll 1$ and $n\varepsilon \ll 1$.

More generally, even without assuming independence, let E_i be the event of an error at step i, and suppose $P(E_i) \leq \varepsilon$ for all $i = 1, \ldots, n$. By the union bound,

$$P\left(\bigcup_{i=1}^{n} E_i\right) \leq \sum_{i=1}^{n} P(E_i) \leq n\varepsilon.$$

This mathematical property implies that as long as we remain in a Softmax framework with **finite logits and no hard mask**, the information path (changing probability values) faces an **insurmountable mathematical ceiling**: invalid branches may be pushed to extremely low probability, but they can never be removed from the support.

In open-domain settings, the reasoning-chain length n can be large and hard to upper-bound in advance. As long as each step retains a non-zero error probability, the bound above grows with n, making the overall failure risk eventually non-negligible.

Key distinction. Engineering control can indeed yield "negligible risk" in **closed-domain, short-chain** scenarios. But for complex logical reasoning that requires **formal guarantees**, the mathematical fact $P > 0$ becomes a hard **ceiling of reliability**.

2.2 Epistemological Tension: Likelihood vs. Necessity

Softmax's property is only the **mathematical basis** of hallucinations. The deeper **epistemological root** is the tension between a statistical training objective and a rational target. Modern LLMs minimize next-token prediction error and maximize the estimated probability of tokens that appear in training data. This **statistical** objective pursues **Likelihood**—that is, **high-probability correlation**—rather than **logical necessity**. What the system learns is "what symbols usually appear under what conditions" (a **statistical shadow of logic**), not constraints of "what must appear."

This yields an **epistemological tension** between likelihood and necessity:

- A **probabilistic mechanism** admits degrees: it allows "highly improbable" but refuses "impossible." Under probability, everything remains possible; only weights differ.
- A **logical mechanism** requires the absolute applicability of excluded middle and non-contradiction. Deductive steps are either valid or invalid; there is no "approximately valid."

Existing improvements remain inside the probabilistic frame: either injecting explicit logical information into prompts and context to change conditional probabilities (e.g., **Logic-of-Thought (LoT; Liu et al., 2025)**), or improving auditability via multi-agent debate (e.g., **Sparse Multi-Agent Debate (SMAD; Yang et al., 2026)**). Yet benchmark evidence (e.g., **LogiConBench (Chen et al., 2026)**) reveals a long tail of logical inconsistency: **as long as invalid choices remain in the support, "suppression" is never "exclusion."** To achieve exclusion, one must enter the **support-level operations** of the structural path in Section 3.

2.3 Civilizational Risk: Criterion Drift and a Self-Reinforcing Loop

In our view, hallucination is not only a technical defect; it is a civilizational-scale epistemic risk. As LLMs become infrastructure for content production and decision-making, the criterion of validity is drifting from "**verifiable**" toward merely "**generable**."

Traditional knowledge ecosystems rely on a **verificationist criterion**—traceable sources and checkable evidence. Generative-AI ecosystems drift toward a **generativist criterion**—it suffices that something "looks true" (plausible) or is probabilistically coherent. This **criterion drift** triggers model collapse (Shumailov et al., 2024): generated content flows back into training data, flattening distribution tails, and models gradually learn to generate content that "looks like what models generate." The system converges toward **model-ness** rather than truth—treating probabilistic quicksand as though it were logical bedrock.

One might appeal to **data filtering**, but that misses the **recursive** nature of criterion erosion: the filtering standard itself may already be eroded by the generativist criterion, like measuring oneself with the same biased ruler. To break this recursion at the mechanism level, the most direct and reliable path is to introduce checkable hard constraints inside the generation mechanism (the structural path), anchoring "looks true" back to "decidable as true" where possible. Addressing civilizational risk therefore requires pushing the spectrum toward the hard end—precisely the structural path developed in Section 4.

3 From Values to Domains: A Continuum of Progressive Hardening

Mainstream attempts often presuppose that logical rules should be **encoded as information**, and that by adjusting inputs one can **change the values** of a probability distribution. However, as long as we remain on the information path with an unchanged support, the probability of invalid outputs can only be suppressed, never driven to exactly zero.

We propose a **continuous spectrum of logical interventions on probability**. One end is the **information path**—changing probability values to **modulate** risk. The other end is the **structural path**—reconstructing feasible sets to **eliminate** risk. Although constraint hardness can appear to increase gradually in practice, achieving "exact zero" reveals a mathematical discontinuity.

3.1 THE CONTINUUM: FROM VALUES TO SUPPORT

A discrete probability distribution is characterized by two dimensions:

- **Values**: the numerical probabilities of events, i.e., the values of $P(t \mid c)$.
- **Support**: the set of outcomes with non-zero probability mass, i.e., $\mathrm{supp}(P(\cdot \mid c)) = \{t \in V \mid P(t \mid c) > 0\}$.

Under the standard Softmax setting with **finite logits and no hard mask**, $\mathrm{supp}(P(\cdot \mid c)) = V$. Logical intervention can proceed along two fundamentally different directions:

Information path (soft end). By modifying context, adjusting logits, or adding biases, we change the **values** of the distribution—making logically valid tokens "heavier" and invalid tokens "lighter." Prompt engineering and logical information injection are typical examples. The defining feature is that the **support remains unchanged** ($\mathrm{supp}(P(\cdot \mid c)) = V$). These methods can drive invalid-token probabilities very low, but because of Softmax's mathematics, $P(t \mid c) > 0$ always holds—"highly improbable" is not "impossible."

From an engineering perspective, constraints may appear to strengthen gradually: stronger prompts, stricter process norms, stronger checking tools, and so on. However, with respect to the goal of **exact prohibition** (probability exactly zero), there is a key threshold: **as long as we stay within standard Softmax over finite logits, an invalid branch remains a non-zero-probability event; once we introduce hard masking to reconstruct the feasible set, invalid branches are explicitly excluded from the possibility space.** We therefore do not deny the practical continuum of engineering implementations; rather, we emphasize that, for "exact zero," there is a **mathematical–mechanistic discontinuity**.

Structural path (hard end). We reconstruct the feasible set by removing logically invalid tokens from the candidate set *before* Softmax sampling. Operationally:

$$P'(t \mid c) = \begin{cases} \dfrac{\exp(z_t)}{\sum_{t' \in V_{\text{valid}}} \exp(z_{t'})}, & t \in V_{\text{valid}} \\ 0, & t \notin V_{\text{valid}} \end{cases}$$

where V_{valid} is the feasible set determined by a validity checker. The crucial change is the normalization domain: the distribution is normalized only over V_{valid}. For any $t \notin V_{\text{valid}}$, the probability is defined as 0. This is not a better approximation; it is a qualitative change in **constraint hardness**: invalid tokens are explicitly excluded from the **possibility space** at the level of reachability, rather than merely assigned low weight.

The mathematical discontinuity between "value operations" and "domain operations" is proved rigorously in the next subsection via Formal Observations 1 and 2.

3.2 FORMAL OBSERVATIONS: THE MATHEMATICAL DISCONTINUITY

Building on the continuum above, we must confront a **formal observation** that blocks all potential detours attempting to realize hard constraints using "cleverer pure-probability tricks":

Formal Observation 1 (the non-zero lower bound of the information path). Consider autoregressive generation based on standard Softmax sampling. Let $|V| \geq 2$ and suppose there exists at least one logically invalid token $t^* \in V$. For any logit-adjustment function $f : \mathbb{R}^{|V|} \to \overline{\mathbb{R}}^{|V|}$, where $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$ denotes the extended real line, the following cases hold:

- **Case 1.** If the output of f is finite (i.e., $f(\mathbf{z}) \in \mathbb{R}^{|V|}$), then after Softmax, $P(t^*) > 0$ holds.
- **Case 2.** If f outputs $-\infty$ for some t^*, then it has effectively removed t^* from the feasible set; the mechanism has already transitioned to the **structural path** (support operation) and is no longer a purely information-path (value-only) operation.
- **Case 3.** If f outputs $+\infty$ for some token (or makes a component diverge), Softmax degenerates—in the limit—into concentrating probability mass on that token (others approach 0). This degeneration likewise steps outside the "finite-logit information path" premise and is best viewed as a structural/degenerate boundary case.

Sketch of proof. When $f(\mathbf{z}) \in \mathbb{R}^{|V|}$ is a finite vector, Softmax yields $\sigma(f(\mathbf{z}))_t \in (0,1)$ for every token t, with $\sum_t \sigma(f(\mathbf{z}))_t = 1$. For **Case 1**, since $e^x > 0$ for any finite real x,

$$\sigma(f(\mathbf{z}))_{t^*} = \frac{e^{f(\mathbf{z})_{t^*}}}{\sum_j e^{f(\mathbf{z})_j}} > 0,$$

hence $P(t^*) > 0$. For **Case 2**, if $f(\mathbf{z})_{t^*} = -\infty$, by convention $e^{-\infty} = 0$, so t^* contributes no probability mass in the denominator, which is equivalent to applying a hard mask that excludes t^* from the feasible set (and thus from the support of the constrained distribution). Therefore, this operation is substantively a feasible-set constraint, not continuous "weight tuning" within $\mathbb{R}^{|V|}$.

Addendum: if a component becomes $+\infty$ (or diverges numerically), Softmax degenerates in the limit to concentrate mass on that component; this is not within the "finite-logit Softmax value operation" premise and is best treated as a structural/degenerate boundary case. □

Formal Observation 2 (the zero-probability guarantee of the structural path). Under support reconfiguration, if a validity checker removes the invalid token from the feasible set (i.e., $V_{\text{valid}} = V \setminus \{t^*\}$), and we define the constrained next-token distribution P' as the distribution normalized only over V_{valid}, then for the invalid token t^*, $P'(t^*) = 0$ holds **exactly**.

Sketch of proof. By definition of the constrained distribution,

$$P'(t) = \begin{cases} \dfrac{e^{z_t}}{\sum_{t' \in V_{\text{valid}}} e^{z_{t'}}}, & t \in V_{\text{valid}} \\ 0, & t \notin V_{\text{valid}} \end{cases}$$

so if $t^* \notin V_{\text{valid}}$, then $P'(t^*) = 0$. □

Synthesis. Observations 1 and 2 sharply characterize the discontinuity between **value operations** (continuous reweighting) and **domain operations** (feasible-set operations). This discontinuity is structural rather than psychological: it does not depend on what we believe, but on how the generation distribution is defined. Concretely, within standard Softmax with **finite logits**, each probability component lies strictly in $(0,1)$ (and sums to 1), so it cannot be exactly 0. Once we introduce hard masking / exclude tokens from V_{valid} (equivalently setting their logits to $-\infty$), we have performed an explicit feasible-set operation and thereby achieved exact zero. Put differently: $-\infty$ is not "a very large negative number"; it is a step outside the "finite real logits" regime.

In engineering, this discontinuity manifests as a hierarchy of realizability: from decidable *syntactic* constraints to broader *structural* constraints (often with higher implementation and computational complexity). See Section 4.3 for the realizability hierarchy within the structural layer.

3.3 Theoretical Roots: Dynamic Logic and Structural Elimination

Our structural path resonates methodologically with **dynamic epistemic logic** (DEL) in the tradition of Johan van Benthem (van Benthem, 2011). In DEL, the normative force of logic comes from **reconstructing the possibility space through information updates**: by "public announcement," a model is restricted to the subset of worlds satisfying φ (with the reachability relation correspondingly restricted), explicitly eliminating worlds that do not satisfy φ.

The **operational kernel** of this paper shares two insights with that tradition. First, **elimination as normative force**: DEL makes non-φ worlds **inaccessible**; our hard mask explicitly excludes invalid tokens from the **feasible set** V_{valid}, so that in the constrained distribution $P'(t) = 0$ whenever $t \notin V_{\text{valid}}$. Both point to the same conclusion: to obtain necessity in the sense of **exact zero / hard guarantees**, one must rely on **structural elimination**, not merely on probabilistic reweighting. Second, **governance of reachability**: DEL treats logic as a reconstruction of reachability relations in the possibility space, while we operationalize it as pruning feasible candidate sets; both elevate logic from "truth-value calculation" to **possibility-space governance**, emphasizing that **inaccessibility** bears normativity more directly than **low probability**.

Honest qualification. We do not claim formal equivalence to DEL. van Benthem studies abstract epistemic agents and modal models; we study neural generation mechanisms and constrained gen-

eration distributions. The former centers on **epistemic accessibility**, the latter on **computational generability**. Yet they are aligned in the shared idea that **hard necessity is implemented via deletion**, rejecting the collapse of logic into a mere statistical shadow of likelihood.

This theoretical anchor provides a **syntax of inquiry** for neurosymbolic research: asking "what structural constraints can implement deletion at the decoding boundary," rather than "how to better reweight probabilities."

4 Three Layers of Logical Intervention: An Operational Framework from Empiricism to Formalism

The spectrum analysis in Section 3 becomes **three threshold regions**, forming a **defense in depth** system from soft to hard. Each layer can be deployed independently (information-layer prompting, normative templates, or structural masking), and layers can also be composed (conflict handling and fallback in Section 4.4). Together they constitute alternative architectures aligned with different epistemological stances.

4.1 Layer 1: Information Layer (Soft End) — Empiricism

Epistemic stance and core mechanism. Empiricism: extract logical propositions via preprocessing, apply rules of inference (e.g., transitivity, contrapositive, double negation) to generate implicit conclusions, and append them to context to improve statistical performance.

Representative practice. Logic-of-Thought (LoT; Liu et al., 2025) injects propositional-logical expansions into prompts; RAG supplies evidence via external knowledge bases.

Mechanism and limit. By changing context, this layer modulates the **values** of conditional probabilities, increasing the probability of valid tokens. Yet logically invalid tokens remain in the **support** and are merely assigned low probability. The information layer provides **statistical enhancement**, but $P(\text{error}) > 0$ remains, and it also faces difficulties such as **knowledge explosion** and **conflict handling**.

4.2 Layer 2: Normative Layer (Intermediate Threshold) — Rationalism

Epistemic stance and core mechanism. Rationalism: constrain thinking through procedural **discipline** and **auditability**. Reasoning chains are required to follow structured templates, including explicit rule annotations (e.g., modus ponens), dual representations (symbolic logic and natural language), and cross-checking (multi-agent debate or self-consistency checks).

Representative practice. Symbolic Chain-of-Thought (SymbCoT; Xu et al., 2024) introduces symbolic expressions and rule verification; **Sparse Multi-Agent Debate (SMAD; Yang et al., 2026)** uses sparse communication to improve robustness.

Mechanism and limit. By enforcing process control, this layer raises the **cost** and increases the **visibility** of invalid reasoning, making it easier to detect after the fact. However, the normative layer does not reconstruct feasible sets at the level of defining the next-step distribution (it constrains the process and its auditability). It also faces a fundamental gap: **"claiming compliance" is not "actual compliance."** A model can claim to apply inference rules while producing conclusions that violate the intended logical form, leaving **residual risk**.

4.3 Layer 3: Structural Layer (Hard End) — Formalism

Epistemic stance. Formalism: logic is not information flowing through the system, nor merely discipline over the process; it is a deterministic structure that **reconstructs the possibility space itself**.

Core mechanism. A **validity checker** makes a binary decision (valid/invalid) and hard-masks invalid tokens *before* Softmax sampling, explicitly excluding them from the support. Operationally:

$$P'(t \mid c) = \begin{cases} \dfrac{\exp(z_t)}{\sum_{t' \in V_{\text{valid}}} \exp(z_{t'})}, & t \in V_{\text{valid}} \\ 0, & t \notin V_{\text{valid}} \end{cases}$$

Here V_{valid} is determined by independently specified formal rules via the validity checker; it is **not produced by the model's probability assignment itself**. The key change is the normalization domain: the distribution is normalized only over V_{valid}; for $t \notin V_{\text{valid}}$, probability is defined as 0; for $t \in V_{\text{valid}}$, renormalization sums only over valid candidates.

4.3.1 Interface with the normative layer (not a dependency)

The structural layer's constraint rules are **defined independently**. They can be hard-coded from external formal specifications (e.g., JSON Schema, rule bases), and they may also consume the **checkable intermediate structures** produced by the normative layer—reducing compilation complexity—but this is not a required dependency (the structural layer can operate independently).

Within the structural layer there is a **hierarchy of realizability**. The practical transition strategy is not to cover open domains at once, but to first obtain "fragment-level exact zeros" in decidable subdomains, and then expand hard-constraint coverage modularly.

> **Case A: Syntactic hard constraints (already deployed).** The system maintains a deterministic finite-state machine (FSM) or a context-free grammar (CFG). During decoding, any token that would make the structure invalid (e.g., JSON Schema violation, unbalanced brackets) is **hard-excluded** (probability set to 0). This confirms the engineering feasibility of "changing the feasible set," but it is a syntactic subset of the structural layer.
>
> **Case B: Decidable-fragment constraints (partially validated).** The system maintains a deterministic logical state (e.g., board state, proof state, type environment) and performs **validity checks** on candidate actions. If a candidate causes a state violation (illegal move, type error), it is hard-excluded. In such restricted domains, violations are categorically excluded—transforming "low-probability risk" into a "structural exclusion zone." Ma and Hu (2025) demonstrate feasibility at this level, marking the key transition from "string shape" (syntax) to "decidable semantics."
>
> **Case C: Action-/state-space constraints (frontier).** Model reasoning as state transitions in an action space, maintaining cross-step consistency commitments (e.g., transitive closure, taxonomy consistency, or contradiction shielding). The challenges include state-space explosion, non-monotonic updates, and combinatorial complexity; in open domains, this often exhibits **engineering complexity close to AI-Complete**, and we position it as a research frontier.

As the **last line of defense** in a defense in depth architecture, the structural layer ensures that even if the information layer fails and the normative layer misses issues, logically invalid outputs cannot be instantiated at the token-generation boundary within the coverage of its rules.

Overridability (supplementary). Hard constraints should be explicitly liftable/waivable and should log when they are suspended, so the system can preserve soundness (and communicate uncertainty) when inputs are premise-distorted or undecidable. We provide the full discussion and analogy in Appendix B.

5 Related Work: Positioning on the Information–Structure Spectrum

Our framework recontextualizes existing technical practice. Through the lens of the information–structure spectrum, many engineering practices (e.g., syntactic hard constraints) already occupy the middle ground. Yet the systematic application of **support-level hard structural constraints** to open-domain natural-language reasoning—especially the extension from restricted domains to open-domain action-/state-space constraints—remains largely unexplored, and we view it as a potential **frontier for a new paradigm**.

5.1 Pure Information End (Empirical Path)

Logic-of-Thought (LoT; Liu et al., 2025) and **Hypotheses-to-Theories (HtT; Zhu et al., 2023)** represent the soft end. LoT uses propositional rules (transitivity, contrapositive, double negation) to generate additional logical information and inject it into context; HtT learns reusable textual rule libraries via a two-stage "induction–deduction" process. Both increase logical information to change the **values** of conditional probabilities—making valid outputs "heavier" and invalid ones "lighter." However, these methods keep the **support** unchanged: invalid tokens still participate in Softmax normalization. Their probabilities may be compressed to extremely small values, but they can never be structurally deleted from the support, leaving non-zero residual risk.

5.2 Middle Ground (Norms and Soft Structure)

Symbolic Chain-of-Thought (SymbCoT; Xu et al., 2024) and **Selection-Inference (Creswell et al., 2023)** introduce explicit reasoning steps and formatting discipline, making the process auditable, but still rely on the model's self-enforcement and do not hard-prune candidates at the decoding boundary.

Certified Deductive Reasoning (LogicGuide; Poesia et al., 2024) sits near the boundary between the normative and structural layers. Through *guides* and **Constrained Semantic Decoding (CSD; Poesia et al., 2022)**, it restricts generation within "guided blocks" to sets deemed valid by external reasoning tools. This approach uses shielding/constrained decoding as its core (with soft bias when necessary), so that generation within guided blocks is constrained **by construction** by the valid set determined by external tools, rather than relying on post-hoc rejection sampling.

LogicGuide validates a key fact: when moving from "process discipline" (normative) toward "structural guarantees," constraints must ultimately land on decoding-time control of the candidate set—i.e., "delete unacceptable branches from the generable set," not merely apply soft bias to probability values.

5.3 Structural-Path Pioneers (Hard Constraints, Restricted Domains)

Ma and Hu (2025) implement **hard logit masking** in propositional-logic and chess domains: by setting the logits of invalid tokens to $-\infty$, they explicitly remove them from the support *before* Softmax normalization, achieving probability exactly zero. This demonstrates the feasibility of **support reconfiguration**, but current implementations are largely confined to restricted domains (e.g., legal moves, resolution-proof steps) and have not yet extended to open-domain natural-language semantic constraints.

5.4 Bypass Paths (Non-intervention)

Logic-LM (Pan et al., 2023), **SatLM (Ye et al., 2023)**, and **LINC (Olausson et al., 2023)** take a "bypass path": they **outsource** deduction to external symbolic systems (first-order provers, SAT/SMT solvers), with the LLM serving primarily as a **semantic parser**. Their two-stage architecture is: (i) use Softmax generation to translate natural language into a formal representation; (ii) call a deterministic external solver to execute symbolic deduction, discarding probabilistic generation in the deduction phase.

These approaches obtain logical guarantees during deduction, but **not by constraining a probability distribution**; rather, they replace the probabilistic mechanism during the deduction phase. They are therefore not "constraining probability with logic," but "**replacing probability with logic**."

Key distinction. Such approaches still rely on probabilistic generation in the semantic-parsing phase. Their fragility stems precisely from this: if the LLM mistranslates in stage (i), the stage-(ii) solver will produce a "correct" derivation from false premises—yielding systematic failures. This differs fundamentally from the structural path, which enforces hard constraints at each token-generation boundary.

5.5 Post-hoc Verification (After Generation)

FoVer (Pei et al., 2025) represents post-generation verification: it translates CoT into executable first-order logic, checks reasoning chains via an SMT solver (e.g., Z3), and triggers feedback-based

correction loops. Compared with "blocking during generation" (the structural path), post-hoc verification has a structural weakness: it allows errors to be instantiated first, relying on rollback/correction to absorb them. This increases cost and typically does not guarantee convergence to an error-free output. By contrast, the structural path attempts to prune invalid branches at the **decoding boundary**, making certain errors mechanistically uninstantiable.

5.6 Our Position: The Spectrum Gap

Existing techniques cover a wide range from information injection to syntactic hard constraints. However, the systematic application of **support-level hard structural constraints** to open-domain natural-language reasoning—especially the extension from restricted domains (syntax, propositional logic) to open-domain action-/state-space constraints—remains largely unexplored.

Ma and Hu (2025) validate hard masking in restricted domains, but do not yet address structural constraints based on cross-step consistency commitments and **action spaces** in open domains. We focus on the hardest end of the spectrum: through **decoding-time support reconfiguration**—rather than solver outsourcing (Section 5.4) or post-hoc verification (Section 5.5)—approaching logical necessity at the generation boundary and categorically eliminating systemic risks within constraint coverage.

6 Conclusion: Boundaries, an Evolution Roadmap, and an Invitation

Note on the extended version. To preserve the 10-page main-text limit while keeping the full argument, we place the original Sections 6.1–6.2 (summary and roadmap) in Appendix A.

Closing: A Research Agenda as Invitation. This paper offers a **diagnosis** (why hallucinations have structural inevitability), a **map** (the soft-to-hard continuum and three-layer stack), and a **direction** (why we must move toward the structural endpoint), rather than a fully sufficient engineering implementation.

This diagnosis touches a century-long debate in AI: pure probabilistic paths have hit a **principled wall** for logical reliability—Softmax's mathematics cannot carry necessity; it can only approximate likelihood. The Values→Support continuum provides a **reconciliation map**: it delineates the sovereign territory of probabilistic mechanisms and the necessary intervention points of formal logic, transforming an either–or replacement relationship into a division of labor at the decoding boundary. This marks a historic shift from **soft alignment** toward **hard guarantees**.

Choosing formal systems, optimizing real-time constraint checking, integrating constraints with neural architectures, and devising decidable approximations in open domains require deep collaboration among formal logicians, neural-network engineers, and cognitive scientists. We offer this paper as an invitation to evolve "constraining probability with logic" from probabilistic-tuning techniques into a system architecture for **possibility-space governance**, rebuilding structural anchors for verifiability in the era of generative AI.

We humbly acknowledge probability's sovereignty in pattern recognition and open-ended generation, and we honestly face the limits posed by undecidable domains. Yet within formalizable fragments—at the boundaries where necessity is required—the structural path has an irreducible normative priority. This is not a technical preference; it is mathematical necessity.

Because **the foundations of civilization cannot be built on probabilistic quicksand**.

References

References

Chen, Z., Zhou, C., Cheng, F., Yip, T. P., Liu, F., Wang, Y., Chai, J., Wang, X., Yin, G., Lin, W., Li, H., Li, B., & Lin, Z. (2026). LogiConBench: Benchmarking Logical Consistencies of LLMs. *International Conference on Learning Representations (ICLR).* (Poster). https://openreview.net/forum?id=ULEHJkolxB

Creswell, A., Shanahan, M., & Higgins, I. (2023). Selection-Inference: Exploiting Large Language Models for Interpretable Logical Reasoning. *International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=3Pf3Wg6o-A4

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems (NeurIPS)*. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

Liu, T., Xu, W., Huang, W., Zeng, Y., Wang, J., Wang, X., Yang, H., & Li, J. (2025). Logic-of-Thought: Injecting Logic into Contexts for Full Reasoning in Large Language Models. *NAACL 2025*. https://aclanthology.org/2025.naacl-long.510/

Ma, F., & Hu, A. J. (2025). Logically Constrained Decoding. *Proceedings of the 3rd Workshop on Mathematical Natural Language Processing (MathNLP 2025)*. https://aclanthology.org/2025.mathnlp-main.11/

Olausson, T. X., Gu, A., Lipkin, B., Zhang, C. E., Solar-Lezama, A., Tenenbaum, J. B., & Levy, R. P. (2023). LINC: A Neurosymbolic Approach for Logical Reasoning by Combining Language Models with First-Order Logic Provers. *EMNLP 2023*. https://aclanthology.org/2023.emnlp-main.313/

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems, 35*, 27730–27744. https://proceedings.neurips.cc/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html

Pan, L., Albalak, A., Wang, X., & Wang, W. (2023). Logic-LM: Empowering Large Language Models with Symbolic Solvers for Faithful Logical Reasoning. *Findings of EMNLP 2023*. https://aclanthology.org/2023.findings-emnlp.248/

Pei, Y., Du, Y., & Jin, X. (2025). FoVer: First-Order Logic Verification for Natural Language Reasoning. *Transactions of the Association for Computational Linguistics, 13*, 1340–1359. https://dblp.org/rec/journals/tacl/PeiDJ25

Poesia, G., Polozov, A., Le, V., Tiwari, A., Soares, G., Meek, C., & Gulwani, S. (2022). Synchromesh: Reliable Code Generation from Pre-trained Language Models. *International Conference on Learning Representations (ICLR)*. (Poster). https://openreview.net/forum?id=KmtVD97J43e

Poesia, G., Gandhi, K., Zelikman, E., & Goodman, N. D. (2024). Certified Deductive Reasoning with Language Models. *Transactions on Machine Learning Research*. https://openreview.net/forum?id=yXnwrs2Tl6

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature, 631*, 755–759. https://doi.org/10.1038/s41586-024-07566-y

van Benthem, J. (2011). *Logical Dynamics of Information and Interaction*. Cambridge University Press. https://www.cambridge.org/core/product/identifier/9780511974533/type/book

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems (NeurIPS)*. https://proceedings.neurips.cc/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html

Xu, J., Fei, H., Pan, L., Liu, Q., Lee, M.-L., & Hsu, W. (2024). Faithful Logical Reasoning via Symbolic Chain-of-Thought. *ACL 2024*. https://aclanthology.org/2024.acl-long.720/

Yang, H., Cheng, F., Yao, T., Chai, J., Wang, X., Yin, G., Lin, W., Yang, M., Wang, Y., Liu, F., Li, H., & Kar, S. (2026). Enhancing Complex Symbolic Logical Reasoning of Large Language Models via Sparse Multi-Agent Debate. *International Conference on Learning Representations (ICLR)*. (Poster). https://openreview.net/forum?id=rdE9qxGfIv

Ye, X., Chen, Q., Dillig, I., & Durrett, G. (2023). SatLM: Satisfiability-Aided Language Models Using Declarative Prompting. *Advances in Neural Information Processing Systems (NeurIPS)*. https://papers.nips.cc/paper_files/paper/2023/hash/8e9c7d4a48bdac81a58f983a64aaf42b-Abstract-Conference.html

Zhu, Z., Xue, Y., Chen, X., Zhou, D., Tang, J., Schuurmans, D., & Dai, H. (2023). Large Language Models can Learn Rules (Hypotheses-to-Theories, HtT). *arXiv*. https://doi.org/10.48550/arXiv.2310.07064

USAGE OF AI

In this work, we made limited use of LLMs as an assistive writing tool. Specifically, we used LLMs to replace synonyms, restructure sentences, and brainstorm alternative ways of expressing ideas within paragraphs. All conceptual contributions, research design, experiments, analyses, and final writing decisions were made by the authors. The authors take full responsibility for the accuracy and originality of the content.

A EXTENDED CONCLUSION DISCUSSION

A.1 SUMMARY OF CORE INSIGHTS

First, the unavoidable **structural tendency** of hallucinations: Softmax-based LLM generation exhibits a **non-zero lower bound**. This is not an accidental defect due to under-training; it is a systemic tendency jointly determined by the mathematical mechanism ($P > 0$) and epistemic presuppositions (statistical correlation vs. logical necessity).

Second, the **values-to-domain spectrum**: logical intervention is not a single method, but a continuum from changing probability values (soft alignment) to reconstructing feasible sets (hard guarantees). The information path faces an insurmountable mathematical ceiling (Formal Observations in Section 3.2); the structural path explicitly excludes invalid tokens via hard masking, achieving probability exactly zero within the constrained fragments. At the system level, risk decreases as constraint coverage increases.

Third, a **three-layer evolution framework**: an operational stack from the information layer (empirical), through the normative layer (rational), to the structural layer (formal), alongside a realizability hierarchy—syntactic hard constraints (deployed), decidable-fragment constraints (partially validated), and action-/state-space constraints (frontier).

A.2 BOUNDARIES AND AN EVOLUTION ROADMAP

Our framework has both epistemic and engineering boundaries. The most important engineering shift is to explicitly mark where constraints apply and where they do not, making this boundary a first-class object in system design (rather than an implicit "try to comply" in prompting).

(1) The decidability boundary. Formal-systems theory suggests that sufficiently expressive systems cannot simultaneously maximize (i) expressiveness, (ii) decidability/termination, and (iii) completeness. The structural layer therefore offers **soundness** but may incur **false negatives** (conservatively labeling "true but unprovable" or "outside the constrained fragment" as invalid). We adopt **soundness over completeness**: better to reject a valid output than to allow a rule-violating one.

(2) The hierarchy of realizability and an evolution roadmap. This is not a prophecy; it is a **research programme**: continuously decompose open-domain reliability problems into decidable subdomains, progressively expand structural-path coverage, and define verifiable success criteria at each step.

Deployed (productionized). Syntactic hard constraints (FSM/CFG, JSON Schema, grammar sampling) are already validated in production. The bottleneck is that they ensure only *formal well-formedness*, not semantic consistency. Formal languages, compiler theory, and type systems are needed to turn key semantic constraints into decidable conditions.

Validated (deployable). Decidable logical fragments (propositional resolution/SAT, typed programming, formal proofs) can use external checkers to hard-prune candidate sets at decoding time. The bottleneck is expanding coverage from restricted to semi-open domains. The success criterion is zero hallucination *within the constrained fragment* (i.e., within the invalidity category defined by the constraints) in verifier-covered settings such as code generation and theorem proving.

Mid-term target (vertical semantic structures). Domain-specific semantic structures (legal/medical ontologies, knowledge graphs, rule bases) compress parts of open semantics into finite semantic spaces, enabling verification at critical sub-tasks. The bottleneck is the cost of ontology coverage and consistency maintenance. The target is verifiable guarantees at key nodes in vertical domains (e.g., statutory citations, contraindications).

Long-term open problem (open-domain action-/state-space constraints). Open-domain action constraints (action spaces, commitment states, neurosymbolic hybrids) often show engineering complexity close to AI-Complete in general. The bottleneck is decidable approximation and computational-complexity control. The target is materially improved reliability for open-domain complex reasoning, with explicit trade-offs among guarantee scope, cost, and residual risk.

This roadmap makes "evolving from soft to hard" an actionable guide: in practice, one can segment a reasoning chain, prioritize structural hard constraints at critical decision nodes, and use normative auditing plus information-layer enhancement elsewhere—forming a defense in depth with different strengths covering different fragments.

B Overridability: Epistemic Closure and Creativity

Hard constraints in the structural layer must be designed to be **overridable**—i.e., explicitly liftable/waivable as constraints (distinct from "coverage" as scope). This is both an engineering robustness requirement and an epistemological necessity for resisting **criterion drift**. If the information layer makes premises explicit incorrectly, or the normative layer provides misleading intermediate structures, the structural layer could derive conclusions that are "valid but absurd" (formally valid yet premise-distorted). Therefore, when facing undecidability or high-confidence conflict, the system must **fall back** to the normative layer for re-auditing, or explicitly return `unknown`.

This embodies an epistemic posture of **soundness over completeness**: when the system cannot reliably decide, it refuses to hide uncertainty behind fluent text, preventing "generability" from being mistaken for "verifiability." Fallback and constraint suspension should be **logged and marked** (metadata carrying an **epistemic label**) so that recipients can allocate trust appropriately.

B.1 Overridability and creativity: an analogy

Critics worry that hard constraints kill creativity. Consider natural-language grammar: by default, "He walks" is grammatical while "The road walks him" is not, yet a poet may **deliberately violate** grammar to create an effect ("The road walks him into dusk"). The key is that violation is conscious, marked, and meaningful only against a background of default rules.

Hard logical constraints are not the enemy of creativity; they provide its **reference frame**. **Overridability** supports a layered epistemic architecture: by default, constraints are strict (ensuring safety/reliability); in creative contexts, specific constraints may be explicitly suspended, while the suspension itself is recorded. This is not a crude trade-off between safety and creativity; it is a **demarcation** that assigns each its rightful territory.