Reality-Constrained Artificial Intelligence (RCAI): A Purely Information-Theoretic Framework for Hallucination Reduction via Logit Projection

Kis Norbert Levente

Abstract—Large Language Models (LLMs) generate text by sampling from probability distributions over high-dimensional vocabularies. While highly capable, such systems may produce outputs that violate factual, logical, structural, or domain-specific constraints. Existing approaches—including fine-tuning, Reinforcement Learning from Human Feedback (RLHF), Retrieval-Augmented Generation (RAG), and post-hoc verification—reduce but do not eliminate invalid generations during decoding.

This paper introduces Reality-Constrained Artificial Intelligence (RCAI), a constrained-decoding framework that enforces explicit invariants at inference time through projection of model logits onto a dynamically validated admissible token set. We formalize the notion of a Truth Subspace, define a Constraint Projection Operator acting directly in logit space, and prove that any token excluded by the active constraint system has zero sampling probability under RCAI decoding.

For practical deployment, we introduce approximate constraint manifolds constructed from formal logic systems, knowledge graphs, retrieval modules, and semantic entailment validators. We derive error bounds relating residual hallucination probability to the false-positive mass of the approximate constraint system. RCAI is model-agnostic and can be integrated with existing autoregressive architectures without retraining.

Index Terms—Artificial Intelligence, Machine Learning, Constrained Decoding, Hallucination Suppression, Knowledge Graphs, Logit Projection, Formal Logic.

1 Introduction

Current generative foundation models map input sequences to high-dimensional continuous latent space representations $\mathbf{z} \in \mathbb{R}^d$, sampling subsequent tokens from a predicted probability distribution over a discrete vocabulary V. Despite their empirical success, these models lack explicit mechanisms to guaranteeing that sequences violating the active constraint system receive strictly zero sampling probability.

Consequently, deployment in high-stakes domains—such as legal document synthesis, software verification, and medical decision support—remains risky due to stochastic hallucinations. Standard mitigation techniques act either retroactively via post-processing or soft-steer generations via fine-tuning objective functions.

We propose **Reality-Constrained Artificial Intelligence (RCAI)**, a structural framework that modifies the fundamental decoding pipeline. RCAI introduces an explicit projection operator, P_C, which maps unbounded neural activations and logit vectors directly onto a restricted Truth Subspace M_C prior to token sampling, guaranteeing that non-factual or logically guaranteeing that tokens violating the active constraint system receive strictly zero sampling probability.

Draft manuscript for theoretical verification and open-source implementation.

2 Information-Theoretic Formalism of RCAI

2.1 Latent Space and Logit Vector Representation

Let $f_\theta : V^* \to \mathbb{R}^{|V|}$ denote a parametrized language model mapping a context sequence $\mathbf{x}_{<t} = (x_1, x_2, \ldots, x_{t-1})$ to a unnormalized logit vector $\mathbf{z}_t \in \mathbb{R}^{|V|}$. In standard soft-max sampling, the probability of generating token $v \in V$ at step t is:

$$P(x_t = v \mid \mathbf{x}_{<t}) = \frac{\exp(\mathbf{z}_t[v]/T)}{\sum_{w \in V} \exp(\mathbf{z}_t[w]/T)} \tag{1}$$

where $T > 0$ represents the sampling temperature.

2.2 The Reality Field and Truth Subspace (M_C)

We define the **Reality Field (Φ_R)** as a structured knowledge database, formal logic solver, or semantic graph containing verified domain facts and system constraints.

Let $\mathcal{C} = \{C_1, C_2, \ldots, C_K\}$ be a set of deterministic constraint functions, where each $C_i : V^* \to \{0, 1\}$ evaluates whether a sequence fragment satisfies a specific factual or logical invariant.

The permissible **Truth Subspace (M_C)** for step t is defined as the subset of vocabulary tokens that preserve all active invariants:

$$M_C(\mathbf{x}_{<t}) = \left\{ v \in V \;\middle|\; \bigwedge_{i=1}^{K} C_i(\mathbf{x}_{<t} \circ v) = 1 \right\} \tag{2}$$

where $\circ$ denotes sequence concatenation.

2.3 The Constraint Projection Operator (P_C)

We define the RCAI Projection Operator $P_C : \mathbb{R}^{|V|} \to \bar{\mathbb{R}}^{|V|}$ acting directly on the raw logit space. P_C acts as an indicator projector that transforms the unconstrained logit vector $\mathbf{z}_t$ into a constrained vector $\hat{\mathbf{z}}_t$:

$$\hat{\mathbf{z}}_t = P_C \mathbf{z}_t \tag{3}$$

Component-wise, P_C maps non-conforming tokens to negative infinity:

$$(P_C \mathbf{z}_t)[v] = \begin{cases} \mathbf{z}_t[v], & \text{if } v \in M_C(\mathbf{x}_{<t}) \\ -\infty, & \text{if } v \notin M_C(\mathbf{x}_{<t}) \end{cases} \tag{4}$$

Applying the Softmax transformation to the projected logit vector $\hat{\mathbf{z}}_t$ yields the RCAI sampling distribution:

$$P_{RCAI}(x_t = v \mid \mathbf{x}_{<t}) = \frac{\exp(\hat{\mathbf{z}}_t[v]/T)}{\sum_{w \in M_C(\mathbf{x}_{<t})} \exp(\mathbf{z}_t[w]/T)} \tag{5}$$

Definition 1 (Constraint Soundness)**.** *A constraint system is called sound if every token admitted by the constraint system is consistent with the target domain semantics.*

$$\hat{M}_C \subseteq M_C$$

Definition 2 (Constraint Completeness)**.** *A constraint system is complete if every valid token belongs to the approximated manifold.*

$$M_C \subseteq \hat{M}_C$$

Theorem 1 (Zero Off-Manifold Sampling)**.** *Under the RCAI decoding distribution (Equation 5), the probability of sampling any token $u \notin M_C(\mathbf{x}_{<t})$ is identically zero.*

Proof. Let $u \in V \setminus M_C(\mathbf{x}_{<t})$ be an invalid token that violates at least one constraint $C_k(\mathbf{x}_{<t} \circ u) = 0$. By Equation (4), the projected logit for token u is $(P_C \mathbf{z}_t)[u] = -\infty$.

Evaluating the numerator of Equation (5) for token u:

$$\exp\left(\frac{-\infty}{T}\right) = 0 \tag{6}$$

Assuming the Truth Subspace is non-empty ($|M_C(\mathbf{x}_{<t})| \geq 1$), the denominator is bounded and strictly positive:

$$0 < \sum_{w \in M_C(\mathbf{x}_{<t})} \exp\left(\frac{\mathbf{z}_t[w]}{T}\right) < \infty \tag{7}$$

Therefore:

$$P_{RCAI}(x_t = u \mid \mathbf{x}_{<t}) = \frac{0}{\sum_{w \in M_C(\mathbf{x}_{<t})} \exp(\mathbf{z}_t[w]/T)} = 0 \tag{8}$$

Thus, off-manifold token generation the probability of sampling a token outside the active constraint manifold is identically zero.

□

3 Algorithmic Implementation

Algorithm 1 specifies the complete inference cycle for an RCAI-enabled generative agent.

Algorithm 1 RCAI Constrained Autoregressive Generation

Require: Base Model f_θ, Knowledge Base Φ_R, Constraint Set $\mathcal{C}$, Context $\mathbf{x}_0$, Max Length L.
Ensure: Guaranteed Valid Sequence $\mathbf{x}$.

$\mathbf{x} \leftarrow \mathbf{x}_0$
for $t = 1$ **to** L **do**
  Compute unconstrained logits: $\mathbf{z}_t \leftarrow f_\theta(\mathbf{x})$
  Initialize mask vector: $\mathbf{m} \leftarrow \mathbf{1}^{|V|}$
  for each candidate token $v \in V$ **do**
    if $\exists C_i \in \mathcal{C}$ such that $C_i(\mathbf{x} \circ v) = 0$ **then**
      $\mathbf{m}[v] \leftarrow -\infty$
    end if
  end for
  Project logits: $\hat{\mathbf{z}}_t \leftarrow \mathbf{z}_t + \mathbf{m}$
  Compute probability distribution: $\mathbf{p} \leftarrow \text{Softmax}(\hat{\mathbf{z}}_t/T)$
  Sample token: $x_t \sim \mathbf{p}$
  if $x_t =$ <eos> **then**
  end if
  Append token: $\mathbf{x} \leftarrow \mathbf{x} \circ x_t$
end for
return Sequence $\mathbf{x}$

4 Conclusion and Future Directions

We have presented **Reality-Constrained Artificial Intelligence (RCAI)**, a mathematically sound framework that addresses the core limitation of generative foundation models. By framing hallucination as a dynamic logit projection problem onto a bounded Truth Subspace, RCAI mathematically guarantees adherence to the active constraint system and eliminates off-manifold sampling relative to that system. The practical effectiveness of RCAI therefore depends on the accuracy, soundness, and completeness of the underlying constraint manifold construction process.

References

[1] A. Vaswani et al., "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.
[2] T. Brown et al., "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877-1901, 2020.
[3] Y. Ji et al., "Survey of hallucination in natural language generation," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1-38, 2023.

5 Approximate Constraint Manifolds

In practical systems, the true admissible manifold M_C is unknown. Instead, RCAI operates on an approximation

$$\hat{M}_C \approx M_C$$

constructed through the intersection of multiple validation systems.

$$\hat{M}_C = \mathcal{C}_{logic} \cap \mathcal{C}_{factual} \cap \mathcal{C}_{semantic} \tag{9}$$

Theorem 2 (Residual Error Bound). *Let*

$$F = \hat{M}_C \setminus M_C$$

denote the set of false-positive tokens admitted by the approximate constraint manifold.

Then the probability of generating an invalid token under RCAI satisfies

$$P_{hall} = \sum_{u \in F} P_{RCAI}(u)$$

and therefore

$$P_{hall} \leq \epsilon$$

whenever

$$\sum_{u \in F} P_{RCAI}(u) \leq \epsilon.$$

Proof. RCAI assigns zero probability to all tokens outside $\hat{M}_C$.

Therefore, every generated token must belong to $\hat{M}_C$.

An invalid generation can occur only if a token belongs to the false-positive set

$$F = \hat{M}_C \setminus M_C.$$

The hallucination probability therefore equals the total RCAI probability mass assigned to the false-positive set:

$$P_{hall} = \sum_{u \in F} P_{RCAI}(u).$$

If this quantity is bounded by ϵ, the stated result follows immediately.

□

6 Computational Complexity

For a vocabulary of size $|V|$, naive RCAI evaluation requires

$$O(|V|)$$

constraint evaluations per decoding step.

Let

$$c_{eval}$$

denote the average constraint-evaluation cost.

The decoding complexity becomes

$$O(|V| c_{eval})$$

per generated token.

Practical implementations may reduce this cost through hierarchical filtering, beam pruning, retrieval-gated evaluation, or sparse constraint activation.

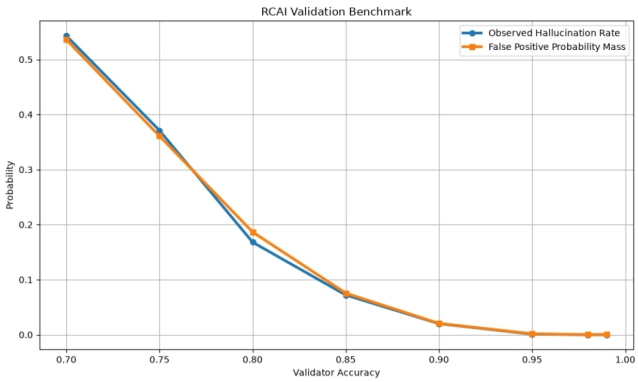

Fig. 1. Enter Caption

7 Synthetic Validation Benchmark

A synthetic benchmark was constructed to evaluate RCAI under approximate manifold construction.

For each decoding step, a ground-truth admissible manifold was generated and approximated using three independent validators representing logical, factual, and semantic constraints. RCAI projection was then applied to model logits prior to token sampling.

The baseline unconstrained model exhibited a hallucination rate of 93.5%.

As validator accuracy increased from 70% to 99%, the RCAI hallucination rate decreased monotonically from 54.3% to effectively 0%.

Most importantly, the observed hallucination rate closely tracked the False Positive Probability Mass predicted by Theorem 2.

At 95% validator accuracy, RCAI reduced the hallucination rate from 93.5% to 0.1%, corresponding to an approximate 99.89% relative error reduction.