# Context-Lifted Predictive States

Representation Complexity, Nested Discourse, and Referential Symmetry

**Saumya Bhandari**

Kageshwori Manohara - 7, Nepal

Theory Version 1.0 – 10 August 2026

**Abstract**

A predictive model of language observes a flat token history, but discourse itself contains nested scenes, clauses, quotations, intentions, ontological frames, and referential bindings. This paper develops a theorem-first framework for asking whether making that latent structure explicit can reduce the *realization cost* of prediction without claiming that it reduces language's intrinsic predictive information. We define an ontology-conditioned nested predictive state (ONPS), distinguish predictive information from model-class-dependent representation cost, and establish six results. First, any deterministic sufficient lift preserves past–future mutual information exactly. Second, the value of context under logarithmic loss is the conditional mutual information $I(Y; C \mid U)$ and decomposes layer by layer across a context stack. Third, a typed nesting language has infinite ordinary Hankel rank, although it has a finite visibly-pushdown realization; this is a known automata-theoretic separation restated as a language-modeling construction. Fourth, a causal finite-alphabet canonicalization removes literal name identity while preserving equality and type relations for a bounded live-register discipline. Fifth, under finite structural resources and visible nesting, the canonicalized ONPS admits a finite weighted visibly-pushdown realization and hence finite nested Hankel rank. Sixth, for clipped logarithmic loss, the approximation error caused by finite entity, ontology, depth, and continuous-state budgets is bounded by measurable overflow tails plus projection error. We also prove an exact decomposition separating structural prediction from residual lexical realization. The result is a conditional mathematical theory and a set of falsifiable complexity hypotheses. It is not a proof that unrestricted English has finite rank, that a Transformer separation has been established, or that large language models are unnecessary. A companion Version 2 protocol is designed to test those empirical claims end to end.

**Central claim.** A causal lift cannot erase intrinsic predictive information. It can, however, change the cost of realizing the same predictive information within a restricted computational model class. The scientific question is whether English exhibits a large enough structured-cost separation to matter in practice.

## 1 Introduction

The string "He got the cup" is locally ambiguous. After a football final, a reader may favor a trophy interpretation; in the ceramics section of a department store, the same local words favor a drinking vessel. The pronoun *he* also has no stable interpretation without a discourse entity state. The local observation is unchanged. The predictive state is not.

This motivates a representation-level hypothesis:

> **Some of the apparent computational difficulty of language prediction may arise because a nested, typed, referential process is observed and modeled as a flat token sequence.**

The hypothesis is deliberately narrower than "language is simple." A sufficient representation cannot remove information that the future actually depends on. What can change is the cost of implementing that sufficient statistic in a chosen model class: the number of states, rank, memory organization, parameters, update operations, and decoder complexity. The distinction is essential. It is also exactly where structured automata provide a clean formal precedent: balanced typed nesting has infinite ordinary Hankel rank and no finite weighted word-automaton realization, yet it has a finite visibly-pushdown realization (Carlyle and Paz 1971; Alur and Madhusudan 2004; Labai and Makowsky 2015).

Discourse theory independently motivates structured context. Grosz and Sidner model attentional state with stack-like focus spaces (Grosz and Sidner 1986); dynamic semantic theories maintain discourse referents across sentences (Kamp and Reyle 1993); frame semantics interprets expressions relative to activated background scenes (Fillmore 1982); and discourse logics allow hierarchical as well as nonlocal attachment (Asher and Lascarides 2003). Predictive-state representations and computational mechanics characterize state by predictions of observables rather than by an inaccessible latent variable (Shalizi and Crutchfield 2001; Boots, Siddiqi, and Gordon 2011; Boots, Gordon, and Gretton 2013). Information bottleneck methods formalize the search for compressed states that retain target-relevant information (Tishby, Pereira, and Bialek 2000; Creutzig, Globerson, and Tishby 2009).

The paper combines these lines into a single object: an *ontology-conditioned nested predictive state* (ONPS). The contribution is not that stacks, ontologies, causal states, register automata, or Hankel matrices are individually new. The contribution is a precise synthesis, a bounded-resource compilation theorem, an approximation envelope, and a clean division between what is proved, what is inherited from prior theory, and what remains an empirical conjecture.

**Contributions.**

1. We distinguish *intrinsic predictive information* from *model-class realization cost* and prove that a sufficient causal lift preserves the former exactly.
2. We formalize context nesting, ontology, and reference in ONPS and derive an exact layerwise information signal.
3. We restate the known ordinary-versus-nested Hankel separation as an exact model-class separation relevant to language modeling.
4. We define a causal canonicalization for names under a bounded live-register discipline and prove invariance under type-preserving renaming.
5. We prove a finite structured realization theorem for the bounded canonicalized model and explicitly exclude the unbounded nominal-pushdown case from the claim.
6. We derive resource-dependent log-loss bounds and an exact structural-versus-lexical uncertainty decomposition.

## 2 Predictive state and two notions of complexity

Let $H_t = (W_1, \ldots, W_t)$ be an observed history and let $F_t$ denote a future target, either the entire future or a finite horizon. A causal representation is a measurable map

$$X_t = \Phi_t(H_t), \qquad \sigma(X_t) \subseteq \sigma(H_t).$$

The filtration condition forbids future leakage by definition.

**Definition 2.1** (Predictive sufficiency). *$X_t$ is predictively sufficient for $F_t$ if*

$$F_t \perp H_t \mid X_t,$$

*or equivalently $\mathbb{P}(F_t \mid H_t) = \mathbb{P}(F_t \mid X_t)$ almost surely.*

**Proposition 2.2** (No-free-information identity). *For every deterministic causal representation $X_t = \Phi_t(H_t)$,*

$$\mathrm{I}(H_t; F_t) = \mathrm{I}(X_t; F_t) + \mathrm{I}(H_t; F_t \mid X_t).$$

*Consequently, $\mathrm{I}(X_t; F_t) \leq \mathrm{I}(H_t; F_t)$, with equality if and only if $X_t$ is predictively sufficient.*

*Proof.* Because $X_t$ is a deterministic function of $H_t$, the pair $(H_t, X_t)$ carries the same information about $F_t$ as $H_t$. Thus

$$\mathrm{I}(H_t; F_t) = \mathrm{I}(H_t, X_t; F_t).$$

The mutual-information chain rule gives

$$\mathrm{I}(H_t, X_t; F_t) = \mathrm{I}(X_t; F_t) + \mathrm{I}(H_t; F_t \mid X_t).$$

The conditional term is nonnegative and vanishes exactly under conditional independence. □

Proposition 2 blocks an overclaim. ONPS does not propose that a coordinate transformation destroys intrinsic past–future information. It proposes that the same information may be easier to realize once nesting, equality, and ontology are explicit.

**Definition 2.3** (Model-class realization cost). *Let $P$ be a prediction problem, $R_P(M)$ the expected risk of a model $M$, $R_P^*$ the Bayes risk, $c(M)$ a declared cost functional, and $\mathcal{M}$ a model class. Define*

$$C_{\mathcal{M}}(P, \varepsilon; c) = \inf \left\{ c(M) : M \in \mathcal{M},\ R_P(M) \leq R_P^* + \varepsilon \right\}.$$

*The cost may include state bits, number of parameters, update time, peak memory, and encoder and decoder cost, provided it is declared before evaluation.*

Two representations can preserve the same predictive information yet yield different $C_{\mathcal{M}}$. The exact separation in Section 5 makes this statement non-metaphorical.

## 3 Ontology-conditioned nested predictive states

**Definition 3.1** (ONPS). *An ontology-conditioned nested predictive state at time $t$ is*

$$X_t = (q_t, \Gamma_t, R_t),$$

*where:*

- $q_t$ *is finite local control or a finite-dimensional predictive coordinate;*
- $\Gamma_t = (\gamma_1, \ldots, \gamma_{d_t})$ *is a stack of typed context frames;*
- $R_t$ *is a live referent register carrying equality, type, role, salience, and finite status features.*

*A frame may contain an ontology type, a discourse purpose, finite local features, and pointers to live referent slots. The visible structural action is one of* CALL*,* RETURN*, or* INTERNAL*.*

The stack is a first model, not a claim that all discourse is perfectly nested. Right-frontier graphs, shared entity stores, or typed hypergraphs may be required for crossing discourse relations. The theorem below asks what follows if a well-nested component can be exposed.

### 3.1 Context stacking and the Ronaldo example

Let $U =$ "He got the cup." Let $Y = (A, S)$ contain the pronoun antecedent $A$ and the sense $S$ of *cup*. In a football discourse, a possible stack is

$$\textsc{World} \triangleright \textsc{Event} \triangleright \textsc{Sport} \triangleright \textsc{Football} \triangleright \textsc{Match},$$

whereas a retail discourse may induce

$$\textsc{World} \triangleright \textsc{Commerce} \triangleright \textsc{Shopping} \triangleright \textsc{DepartmentStore} \triangleright \textsc{Ceramics}.$$

A discourse entity register can retain a distribution over candidates for *he*. The correct formal output is not a forced label but

$$\mathbb{P}(A, S \mid U, X_t).$$

Context changes posterior odds; it does not logically eliminate all unusual interpretations.

## 4 The information-theoretic core

**Theorem 4.1** (Value of context under logarithmic loss). *Let $U$ be a local utterance, $C$ an available context, and $Y$ an interpretation variable. Under logarithmic loss, the reduction in Bayes-optimal expected risk obtained by observing $C$ is*

$$R^*_U - R^*_{U,C} = \mathrm{I}(Y; C \mid U) \geq 0.$$

*Proof.* The Bayes-optimal forecast given $U$ is the true conditional distribution $\mathbb{P}(Y \mid U)$, with risk $\mathrm{H}(Y \mid U)$. Given $(U, C)$, the optimum is $\mathbb{P}(Y \mid U, C)$, with risk $\mathrm{H}(Y \mid U, C)$. Their difference is the definition of conditional mutual information. □

The theorem is an information identity, not evidence by itself that any proposed context encoder is compact or cheap. It tells us what a successful representation must retain.

**Corollary 4.2** (Layerwise context decomposition). *For a context stack $C = (F_1, \ldots, F_d)$,*

$$\mathrm{I}(Y; C \mid U) = \sum_{j=1}^{d} \mathrm{I}(Y; F_j \mid U, F_{<j}).$$

*For any interpretation value $y$ with nonzero conditional probabilities, define*

$$g_j(y) = \log \frac{\mathbb{P}(y \mid U, F_{\leq j})}{\mathbb{P}(y \mid U, F_{<j})}.$$

*Then*

$$\sum_{j=1}^{d} g_j(y) = \log \frac{\mathbb{P}(y \mid U, C)}{\mathbb{P}(y \mid U)}.$$

*Proof.* The first equality is the mutual-information chain rule. The second follows by multiplying the successive probability ratios and cancelling intermediate terms, then taking logarithms. □

The sequence $(g_1(y), \ldots, g_d(y))$ is an exact context-layer information signal. Individual terms may be negative for a particular $y$ even though their expectation is nonnegative.

**Proposition 4.3** (Compression criterion). *For any compressed context $Z = \phi(C)$,*

$$\mathrm{I}(Y; Z \mid U) \leq \mathrm{I}(Y; C \mid U).$$

*Equality holds if and only if $Y \perp C \mid (U, Z)$.*

This is the correct criterion for a "sentence signal": not smoothness, visual elegance, or low dimension alone, but sufficiency for the future and interpretation variables of interest.

## 5 Exact representation separation from nesting

Let $m \geq 2$ and let $c_1, \ldots, c_m$ be call symbols with matching returns $r_1, \ldots, r_m$. Let $D_m$ be the language of properly nested, type-matched words, and let $f = \mathbf{1}_{D_m}$. Its ordinary Hankel matrix is

$$H_f(u, v) = f(uv), \qquad u, v \in \Sigma^*.$$

**Theorem 5.1** (Known flat-rank phenomenon, explicit construction). *For every $k \geq 1$, $H_f$ contains an $m^k \times m^k$ identity submatrix. Hence*

$$\operatorname{rank}(H_f) \geq m^k \quad \textit{for every } k,$$

*and therefore* $\operatorname{rank}(H_f) = \infty$.

*Proof.* For $\boldsymbol{a} = (a_1, \ldots, a_k) \in [m]^k$, let

$$u_{\boldsymbol{a}} = c_{a_1} \cdots c_{a_k}.$$

For $\boldsymbol{b} = (b_1, \ldots, b_k) \in [m]^k$, let

$$v_{\boldsymbol{b}} = r_{b_k} \cdots r_{b_1}.$$

Then $u_{\boldsymbol{a}} v_{\boldsymbol{b}} \in D_m$ exactly when $\boldsymbol{a} = \boldsymbol{b}$. The submatrix indexed by these prefixes and suffixes is $I_{m^k}$. □

This is a direct instance of the standard observation that balanced languages can have infinite ordinary Hankel rank despite finite visibly-pushdown realizations (Labai and Makowsky 2015). It is reproduced because the identity-submatrix proof makes the representation dependence transparent.

**Corollary 5.2** (Exact model-class separation). *For exact realization of $f$ over $\mathbb{R}$, the finite weighted-word-automaton cost is infinite, whereas the visibly-pushdown cost is finite:*

$$C_{\mathrm{WFA}}(f, 0) = \infty, \qquad C_{\mathrm{VPA}}(f, 0) < \infty,$$

*when cost is finite exactly for finite-state/finite-stack-alphabet machines.*

*Proof.* By the Carlyle–Paz characterization, a real-valued word function has a finite weighted word-automaton realization if and only if its ordinary Hankel matrix has finite rank (Carlyle and Paz 1971). Theorem 5 rules this out. A finite visibly-pushdown automaton recognizes $D_m$ by pushing the call type and checking it on return (Alur and Madhusudan 2004). □

Labai and Makowsky prove that a real- or complex-valued well-nested word function is recognized by an $n$-state weighted visibly-pushdown automaton if and only if its nested Hankel matrix has rank at most $n^2$ (Labai and Makowsky 2015). The result does not say that every finite-rank realization is small in practice, but it supplies an exact structured realization notion.

**Lemma 5.3** (Bounded-window versus stack memory). *For every suffix length $L$ and $m \geq 2$, there is a typed-nesting next-symbol prediction problem on which every predictor measurable only from the last $L$ observations has accuracy at most $1/m$, while a stack storing one unmatched call type has accuracy 1.*

*Proof.* Choose $I$ uniformly from $[m]$ and present the history $c_I x^L$, where $x$ is an internal filler symbol. The last $L$ observations are identical for all values of $I$, so no suffix-only predictor can distinguish the correct return $r_I$. Its best accuracy is the largest prior mass, $1/m$. A stack stores $I$ when $c_I$ arrives and predicts $r_I$ exactly. □

*Remark 5.4* (No Transformer separation). Lemma 5 separates bounded suffix memory from explicit stack memory. It does not prove a separation from attention, recurrent networks, structured state-space models, or Transformers. Neural stack and stack-attention architectures already demonstrate that explicit structural memory can be combined with neural sequence models (DuSell and Chiang 2020; Li et al. 2024). The empirical question is comparative cost, not representability in the abstract.

## 6 Referential symmetry and causal canonicalization

Literal names come from a large or open-ended alphabet, but many structural predictions depend only on type and equality: whether the present mention is the same entity as a live referent, not whether its surface form is *Ronaldo*. Register and nominal automata provide established foundations for computation over such infinite alphabets (Kaminski and Francez 1994; Schroeder et al. 2016; Bojanczyk et al. 2021).

Let $\mathcal{T}$ be a finite type set. For each $\tau \in \mathcal{T}$, let $\mathcal{N}_\tau$ be a countably infinite set of names of type $\tau$, and assume the sets are disjoint. Let

$$\mathcal{G} = \prod_{\tau \in \mathcal{T}} \mathrm{Sym}(\mathcal{N}_\tau)$$

be the group of type-preserving name permutations.

A $B$-register configuration consists of at most $B$ occupied slots. Each occupied slot carries a name, its type, finite attributes, and a pin bit. A stack frame may contain only finite frame/ontology attributes and pointers to register slots. A referenced slot is pinned until the frame that references it returns; only unpinned slots may be overwritten. Thus $B$ bounds all names that remain structurally live anywhere in the current state.

**Definition 6.1** (Equivariance). *A transition kernel $T$ and structural output map $O$ are type-equivariant if, for all $g \in \mathcal{G}$,*

$$T(g \cdot x, g \cdot a) = g \cdot T(x, a), \qquad O(g \cdot x) = O(x).$$

*Name-valued outputs may be equivariant rather than invariant; the present theorem concerns structural predictions modulo renaming.*

### 6.1 Finite canonical observations

Given a current register state $R$ and a name token $n \in \mathcal{N}_\tau$, define the causal code

$$\kappa_R(n) = \begin{cases} \mathrm{EQ}_{[i]}, & n \text{ is equal to the name in occupied slot } i, \\ \mathrm{NEW}_\tau, & n \text{ is unequal to every live register name.} \end{cases}$$

Here $[i]$ denotes the equality block of slot $i$ if duplicate slot values are permitted. Non-name observations retain their finite labels. Allocation and eviction actions are included in the canonical transition symbol and are determined from current finite control and the code.

**Theorem 6.2** (Causal name-canonicalization). *Consider two raw histories related by a type-preserving permutation $g \in \mathcal{G}$, initialized from correspondingly permuted register states. Under an equivariant update rule and the same finite allocation discipline, their canonical observation traces are identical, and their states remain in the same $\mathcal{G}$-orbit at every time. Conversely, the canonical trace and initial orbit determine the orbit sequence, though not the literal names.*

*Proof.* Proceed by induction on time. The claim is true initially. Suppose the two current states are related by $g$. Equality of the next name to each live register is preserved by a bijection, and type is preserved by definition of $\mathcal{G}$; therefore the next canonical codes agree. The finite allocation action is the same because it is a function of finite control and the canonical code. Equivariance of the update maps the next states into corresponding orbits. The converse follows because the canonical transition rule is deterministic on orbits. Literal names are intentionally discarded. □

The theorem states exact invariance under name permutation. It also states the limitation: once a name is evicted from every live register, its later reappearance is treated as new. Persistent identity across unbounded inactive history is outside this bounded semantics.

### 6.2 Finite orbit count

Let $A$ be the number of finite attribute combinations per occupied slot. The number of canonical register orbits is finite. A convenient crude upper bound is

$$N_{B,\mathcal{T},A} \leq 4^B \,\mathrm{Bell}(B)(|\mathcal{T}|A)^B.$$

The factors overcount occupancy, pin status, equality partitions, types, and attributes; tightness is not needed.

**Theorem 6.3** (Bounded ONPS structured realization). *Assume:*

1. *finite local control $Q$ and finite discrete predictive coordinate set $Z$;*
2. *finite ontology, frame, role, status, and discourse-operation alphabets;*
3. *a visible call/return/internal partition after the causal context lift;*
4. *the $B$-register pinned discipline above, with every stack pointer targeting a live slot;*
5. *type-equivariant transitions and structural outputs; and*
6. *finite transition/output weights over $\mathbb{R}$ or $\mathbb{C}$.*

*After causal name-canonicalization, the structural word function is recognized by a finite weighted visibly-pushdown automaton. It may be chosen with at most*

$$n \le |Q|\,|Z|\,N_{B,\mathcal{T},A}$$

*canonical control states and a finite stack alphabet. Consequently, its nested Hankel rank is at most* $n^2$.

*Proof.* By Theorem 6.1, raw name identity can be replaced online by a finite code consisting of equality-to-live-slot or a typed-new symbol. Register configurations modulo $\mathcal{G}$ have at most $N_{B,\mathcal{T},A}$ orbits. The product with finite $Q$ and $Z$ is therefore finite. A stack symbol contains only finite frame/ontology/status labels and finitely many pointers in $\{1,\dots,B\}$, so the stack alphabet is finite. The action partition is visible by assumption. The orbit-level weighted transition maps therefore define a finite weighted visibly-pushdown automaton. The nested-Hankel bound follows from Labai and Makowsky's characterization (Labai and Makowsky 2015). □

*Limitation 6.4* (The theorem is conditional). Theorem 6.2 is a sufficiency theorem. It does not establish that English has a finite ontology, bounded live reference, perfectly visible nesting, or a finite useful predictive coordinate. Those are approximation parameters to be measured, not linguistic facts inferred from the theorem.

*Limitation 6.5* (Unbounded persistent names). If arbitrarily many distinct names remain recoverable through an unbounded stack, the system is not compiled by the finite classical VPA construction above. It belongs to nominal/register-pushdown territory, where decidability and equivalence behavior can be substantially harder; some combinations of registers and pushdown storage lead to undecidability results (Murawski, Ramsay, and Tzevelekos 2020). A weighted nominal nested-Hankel realization theorem would be a separate theoretical contribution and is left open.

# 7 Approximation theory for finite resources

Exact finiteness is the wrong empirical question. Let an ideal causal state use arbitrary resources, and let

$$X_t^{B,K,D,R}$$

be an approximation with active-entity budget $B$, ontology budget $K$, explicit depth budget $D$, and continuous predictive dimension $R$.

Let $\Omega_t^{B,K,D}$ be the event that at least one discrete resource overflows: more than $B$ live referents are required, context depth exceeds $D$, or the required ontology item is outside the retained $K$-item ontology. Let $p_t^*$ be an oracle predictive distribution and $q_t^{B,K,D,R}$ the bounded predictor. Define clipped log loss

$$\ell_\alpha(q,y) = -\log \max\{q(y),\alpha\}, \qquad 0<\alpha<1.$$

Its range is at most $L_\alpha = \log(1/\alpha)$.

**Theorem 7.1** (Resource-overflow risk envelope). *Suppose* $q_t^{B,K,D,R} = p_t^*$ *whenever* $\Omega_t^{B,K,D}$ *does not occur and the continuous state is represented exactly. Then*

$$R_\alpha(q^{B,K,D,R}) - R_\alpha(p^*) \le L_\alpha \mathbb{P}(\Omega_t^{B,K,D}).$$

*By the union bound,*

$$R_\alpha(q) - R_\alpha(p^*) \le L_\alpha\big[\mathbb{P}(N_{\text{live}} > B) + \mathbb{P}(d_t > D) + \mathbb{P}(o_t \notin \mathcal{O}_K)\big].$$

*If, outside overflow, the R-dimensional approximation has expected total variation at most $\eta_R$, then*

$$R_\alpha(q) - R_\alpha(p^*) \le L_\alpha \mathbb{P}(\Omega_t^{B,K,D}) + \frac{2\eta_R}{\alpha}.$$

*Proof.* When no overflow occurs and the state is exact, the clipped losses agree. On overflow, the excess loss is at most the full range $L_\alpha$. Taking expectations gives the first inequality; the second follows from the union bound. For the approximation term, the function $x \mapsto -\log \max\{x, \alpha\}$ is $1/\alpha$-Lipschitz on $[0, 1]$. Therefore the pointwise loss difference is at most $\|p_t^* - q_t\|_1/\alpha = 2\,\mathrm{TV}(p_t^*, q_t)/\alpha$. Take expectations and add the overflow term. □

The theorem converts hidden assumptions into measurable curves. Rather than assert "bounded active referents," Version 2 estimates $\mathbb{P}(N_{\text{live}} > B)$ for increasing $B$. Rather than assert "finite ontology," it reports residual risk as $K$ grows. A useful theory predicts favorable tails and diminishing error at modest cost.

## 8 The lexical-realization barrier

Structural state compression does not automatically yield cheap next-token generation. Let $W$ be the next token and let $S = g(W)$ be a deterministic structural or semantic coarsening, such as part of speech, ontology class, referent role, or semantic event type.

**Proposition 8.1** (Structural and lexical uncertainty decomposition). *For every history $H$,*

$$\mathrm{H}(W \mid H) = \mathrm{H}(S \mid H) + \mathrm{H}(W \mid S, H).$$

*Proof.* Because $S$ is a deterministic function of $W$, $\mathrm{H}(S \mid W, H) = 0$. The chain rule gives

$$\mathrm{H}(W, S \mid H) = \mathrm{H}(S \mid H) + \mathrm{H}(W \mid S, H)$$

and also $\mathrm{H}(W, S \mid H) = \mathrm{H}(W \mid H)$. □

The first term can collapse while the second remains large. Therefore a theorem about nested structural realization is not yet a theorem about low-perplexity lexical generation. Version 2 must report both terms and include encoder and decoder cost in every claim.

## 9 The formal research hypotheses

The theory separates three logically independent hypotheses.

**Conjecture 9.1** (Nested-state hypothesis, H1). *A measurable fraction of long-range predictive dependence in English is represented more economically by explicit nested state than by matched flat-history models.*

**Conjecture 9.2** (Referential-symmetry hypothesis, H2). *A measurable fraction of entity-related prediction is invariant under type-preserving renaming and is captured by bounded equality/binding state rather than literal name identity.*

**Conjecture 9.3** (Complexity-collapse hypothesis, H3). *At matched held-out predictive risk, the complete causal ONPS system—encoder, state update, and decoder—achieves a strictly better state/compute Pareto frontier than matched flat neural sequence baselines.*

H1 and H2 may hold while H3 fails. Structural prediction may simplify while lexical realization dominates. A favorable rank diagnostic may fail to translate into latency. These outcomes are scientifically distinct.

## 10 What is new, what is inherited, and what remains open

| **Component** | Status | Role in this paper |
|---|---|---|
| **Ordinary Hankel rank for balanced nesting** | Known phenomenon | Explicit identity-submatrix construction and model-class interpretation |
| **Weighted VPA nested-Hankel theorem** | Prior theorem | Supplies the structured realization criterion |
| **Context value** $\mathrm{I}(Y; C \mid U)$ | Information identity | Gives an operational definition of context value |
| **Causal states / predictive sufficiency** | Prior theory | Defines the target an ONPS approximation seeks |
| **Register and nominal symmetry** | Prior theory | Supplies the language for removing arbitrary names |
| **ONPS synthesis** | Proposed framework | Couples nested context, ontology, reference, and predictive state |
| **Bounded causal canonicalization and risk envelope** | Derived here | Makes symmetry causal and assumptions measurable |
| **Unbounded nominal nested-Hankel theory** | Open | Needed for persistent arbitrary names without a global live budget |
| **Large English complexity collapse** | Empirical conjecture | Tested only in companion Version 2 |

This paper does not claim a general new theorem in classical visibly-pushdown automata. Its strongest formal novelty is the way bounded referential symmetry, causal compilation, structured realization cost, and approximation error are assembled around a language-modeling question. A complete novelty claim for the unbounded nominal case would require deeper comparison with nominal pushdown and weighted-register literature.

## 11 Consequences and non-consequences

The formal results imply:

- realization complexity is model-class dependent;
- flattening a nested process can produce infinite ordinary rank despite finite structured realization;
- context layers have exact, additive information contributions;
- bounded live name identity can be quotiented by type-preserving symmetry in a causal way;
- resource assumptions can be converted to measurable approximation errors.

They do not imply:

- that English has finite nested Hankel rank;
- that finite rank means small parameter count or cheap inference;
- that a stack is the complete structure of discourse;
- that ONPS can be inferred cheaply from raw text;
- that exact lexical generation is low entropy;
- that Transformers cannot represent nesting;
- that current LLMs can be bypassed.

The disruptive claim becomes defensible only if Version 2 finds a leftward shift in the end-to-end predictive-performance-versus-cost frontier, under strict causality, matched baselines, adversarial invariance controls, and an untouched final test set.

## 12 Discussion

The central conceptual move is to treat language history as an observation of a structured predictive process, not to declare sentences literal mechanical systems. The mathematical precedent is strong: state dimension, rank, and memory can change dramatically when a representation exposes the algebraic composition law that the flat sequence hides. The Dyck construction is an exact example. It proves possibility, not prevalence.

The most important object is not a waveform but a causal quotient. Histories that induce the same future-and-interpretation distribution should be identified. ONPS proposes that a useful approximation to that quotient factors into nested frames, ontology, referential equality, and a small residual coordinate. This is falsifiable: if raw history still provides large predictive gains after conditioning on a cheap ONPS state, the proposed factorization is inadequate.

The bounded theorem is intentionally restrictive. It establishes a clean base class from which richer theories can grow. There are at least three extensions: graph-structured discourse beyond a stack; nominal pushdown state with persistent unbounded names; and continuous predictive coordinates learned from observable moments. Each extension should preserve the discipline used here: define the representation, define its composition law, derive the appropriate realization object, and state what empirical tail or approximation error replaces exact finiteness.

## 13 Conclusion

A causal representation cannot make genuine predictive information disappear. It can make that information less expensive to realize. Nested-word theory supplies an exact case in which a flat finite-state representation is impossible while a finite structured realization exists. Discourse theory suggests that language contains analogous nested and referential organization. ONPS turns that suggestion into a bounded mathematical object, derives a causal name-invariant compilation, and replaces hidden finiteness assumptions with measurable resource-error curves.

The theoretical claim of Version 1 is therefore precise:

**If a linguistic process admits a causal, bounded, ontology-conditioned nested predictive state, then its structural prediction can have finite nested realization complexity even when its flat ordinary realization complexity is unbounded; the approximation error of finite resource budgets is controlled by observable overflow and projection terms.**

Whether English occupies that favorable regime, and whether the end-to-end saving survives encoder and lexical-decoder costs, is the subject of Version 2. A positive result would challenge current assumptions about how language prediction should be implemented. A negative result would delimit exactly how far nested contextual structure can simplify the problem.

# A A more explicit canonical-state bound

Suppose each of $B$ slots may be empty or occupied, each occupied slot has one of $|\mathcal{T}|$ types and $A$ attribute values, and occupied slots may share names. The orbit of a register state under type-preserving permutations is determined by:

1. the occupied subset of slots;
2. a partition of occupied slots into equality classes;
3. a type and attribute value for each occupied slot, consistent within an equality class;
4. a pin bit for each occupied slot.

A loose upper bound is

$$N_{B,\mathcal{T},A} \leq 2^B \cdot \mathrm{Bell}(B) \cdot (|\mathcal{T}|A)^B \cdot 2^B = 4^B \,\mathrm{Bell}(B)(|\mathcal{T}|A)^B.$$

If duplicate slot values are forbidden, the Bell factor disappears. If only occupied slots are typed, the bound can be tightened by summing over occupancy $b$.

# B The companion empirical quantity

For a local baseline $L_t$, full raw history $H_t$, and state $X_t$, an operational estimator-relative information-capture ratio is

$$\widehat{\rho} = \frac{\mathrm{CE}(F_t \mid L_t) - \mathrm{CE}(F_t \mid X_t, L_t)}{\mathrm{CE}(F_t \mid L_t) - \mathrm{CE}(F_t \mid H_t, L_t)}.$$

This equals a mutual-information ratio only when the conditional models attain Bayes risk. Version 2 therefore reports it as an operational metric, not as literal information without qualification.

A complementary residual-history gain is

$$\Delta_{\mathrm{hist}} = \mathrm{CE}(F_t \mid X_t) - \mathrm{CE}(F_t \mid X_t, H_t).$$

A small value is necessary but not sufficient for predictive sufficiency; it can also result from an underpowered history model. Hence the protocol uses matched linear and nonlinear baselines and reports complete Pareto frontiers.

## C Version 2 preregistration summary

The companion computational version is frozen around the following requirements:

1. every state is strictly online: $X_t = \Phi(W_{\leq t})$;
2. all encoder, state-update, and decoder costs are included;
3. resource curves vary $(B, K, D, R)$ rather than asserting finiteness;
4. raw, neural, structured, hybrid, shuffled, and random-feature baselines are capacity matched;
5. name permutation, context shuffling, lexical substitution, and nesting destruction test invariance;
6. structural and lexical targets are reported separately;
7. analysis, held-out, and final-test books are split by primary author;
8. the final-test split is opened once after code and metrics are frozen.

## References

Barbara J. Grosz and Candace L. Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204, 1986. https://aclanthology.org/J86-3001/.

Hans Kamp and Uwe Reyle. *From Discourse to Logic*. Kluwer Academic Publishers, 1993.

Charles J. Fillmore. Frame semantics. In *Linguistics in the Morning Calm*, pages 111–137. Hanshin Publishing, 1982.

Nicholas Asher and Alex Lascarides. *Logics of Conversation*. Cambridge University Press, 2003.

Rajeev Alur and P. Madhusudan. Visibly pushdown languages. In *Proceedings of STOC*, pages 202–211, 2004. doi:10.1145/1007352.1007390.

J. W. Carlyle and Azaria Paz. Realizations by stochastic finite automata. *Journal of Computer and System Sciences*, 5(1):26–40, 1971. doi:10.1016/S0022-0000(71)80017-X.

Nadia Labai and Johann A. Makowsky. Hankel matrices for weighted visibly pushdown automata. arXiv:1512.02430, 2015. https://arxiv.org/abs/1512.02430.

Cosma Rohilla Shalizi and James P. Crutchfield. Computational mechanics: Pattern and prediction, structure and simplicity. *Journal of Statistical Physics*, 104:817–879, 2001. doi:10.1023/A:1010388907793.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. arXiv:physics/0004057, 2000.

Felix Creutzig, Amir Globerson, and Naftali Tishby. Past-future information bottleneck in dynamical systems. *Physical Review E*, 79:041925, 2009.

Byron Boots, Sajid M. Siddiqi, and Geoffrey J. Gordon. Closing the learning-planning loop with predictive state representations. In *Robotics: Science and Systems*, 2011.

Byron Boots, Geoffrey J. Gordon, and Arthur Gretton. Hilbert space embeddings of predictive state representations. arXiv:1309.6819, 2013.

Michael Kaminski and Nissim Francez. Finite-memory automata. *Theoretical Computer Science*, 134(2):329–363, 1994.

Lutz Schroeder, Dexter Kozen, Stefan Milius, and Thorsten Wissmann. Nominal automata with name binding. arXiv:1603.01455, 2016.

Mikolaj Bojanczyk, Joanna Fijalkow, Bartek Klin, and Joshua Moerman. Orbit-finite-dimensional vector spaces and weighted register automata. arXiv:2104.02438, 2021.

Andrzej S. Murawski, Steven J. Ramsay, and Nikos Tzevelekos. Bisimilarity in fresh-register automata. arXiv:2005.06411, 2020.

Brian DuSell and David Chiang. Learning context-free languages with nondeterministic stack RNNs. In *Proceedings of CoNLL*, pages 507–519, 2020.

Jiaoda Li, Jennifer White, Mrinmaya Sachan, and Ryan Cotterell. A Transformer with stack attention. In *Findings of NAACL*, pages 4318–4335, 2024.