A Blueprint for Language-Native World Models

Niel Ok
Stanford University
nielok@stanford.edu

April 2025

Abstract

We propose a theoretical framework for *language-native world modeling*, in which structured latent state trajectories—rather than token sequences—serve as the primary object of prediction and control. Extending Shannon's notion of entropy minimization beyond next-token prediction, we introduce an architecture that interprets natural language inputs into semantically structured latent states, models their evolution under hypothetical actions, and optionally verifies their alignment with target goals.

Our system comprises three modules: (1) a *semantic encoder* that maps natural language descriptions of environments and goals into interpretable latent representations z_t; (2) a *latent dynamics model* f_{dyn} that simulates world-state transitions $z_{t+1} = f_{\text{dyn}}(z_t, a_t)$ within a structured or learned embedding space; and (3) a *verifier* that scores the predicted trajectory for goal consistency. To balance interpretability with predictive flexibility, we allow latent states to be projected into learned representations $z'_t = f_{\text{proj}}(z_t)$, and decoded back into structured form $\tilde{z}_t = f_{\text{struct-dec}}(z'_t)$ for round-trip semantic analysis.

Unlike conventional prompting-based LLM use, our formulation separates semantic grounding from trajectory prediction. Simulation is performed entirely in latent space, rather than token space, enabling efficient rollout and planning. This design generalizes model-based reasoning to natural language settings by treating linguistic descriptions as interfaces for initializing and supervising latent world dynamics.

Our work reframes language modeling as a form of predictive compression over evolving semantic configurations, offering a pathway toward modular, interpretable, and language-grounded reasoning systems that simulate not strings, but worlds. This formulation provides a foundation for language-conditioned planning and reasoning agents whose internal simulation dynamics are disentangled from surface-level linguistic form.

1 Background and Related Work

1.1 Shannon Entropy and Predictive Compression

Claude Shannon's foundational work on information theory [8] established that the entropy of a signal source quantifies the theoretical limit of lossless compression. In modern machine learning, this principle underlies the training of language models via next-token prediction: by minimizing cross-entropy loss, models reduce the average number of bits required to represent a sequence, thereby approximating the source distribution. Autoregressive transformers such as GPT [5, 1] and its successors operationalize this principle, achieving remarkable fluency and generalization in textual domains.

1.2 World Models in Reinforcement Learning

In parallel, research in model-based reinforcement learning (MBRL) has explored learning compact representations of the environment to support planning and decision-making. Systems such as PlaNet [3], Dreamer [2], and MuZero [7] learn latent dynamics models that predict future states or rewards, enabling agents to simulate and evaluate hypothetical futures. These models are grounded in continuous sensory spaces (e.g., pixels, proprioception), where the notion of a "world state" is inherently spatial and physical.

1.3 Language Models as Simulators

Recent work has begun to explore the use of large language models as simulators of cognitive and physical processes. Studies on chain-of-thought prompting [9], tool use [6], and agent behavior modeling [4] suggest that LLMs can serve as engines of semantic simulation—predicting outcomes, generating reasoning steps, and even simulating multi-agent interactions. However, these approaches often remain tethered to surface-level token prediction, with no explicit separation between world representation and world evolution.

1.4 Our Contribution: Language-Native World Modeling

Our work introduces a key conceptual shift: we propose that next-token prediction is a special case of a broader class of *world-state prediction* problems. Rather than forecasting the next word in a sentence, we aim to model the evolution of semantically meaningful latent world states as described by natural language. This reframing extends Shannon's theory of entropy minimization from symbol streams to structured trajectories over latent states, and positions LLMs as *semantic world translators*—models that transform language into internal representations of evolving environments.

In contrast to traditional world models grounded in sensorimotor input, our approach remains fully language-native. It operates over semantically structured latent representations derived from text, enabling modularity, interpretability, and compositionality. Crucially, we decouple the processes of *semantic grounding* (mapping text to latent state) and *world evolution* (predicting transitions in latent space), in contrast to conventional prompting pipelines where reasoning and generation are tightly entangled. This separation mirrors classical simulation architectures and allows for more transparent and flexible inference.

By situating our approach at the intersection of information theory, cognitive modeling, and language systems, we outline a framework for agents that simulate, plan, and reason within a latent semantic substrate—without relying on external environments or symbolic logic. In the next section, we formalize this framework and describe how it enables structured prediction, planning, and evaluation directly within latent space.

2 Language-Native World Modeling Framework

We propose a framework that shifts from next-token prediction to structured world-state prediction, enabling language models to simulate and reason about environments described in natural language. Our system interprets natural language observations and goals, translates them into latent representations, and predicts how the world might evolve under hypothetical actions.

2.1 Problem Setup

Let the input be a natural language description of an environment E_t at time t, along with an agent goal G. The objective is to predict a plausible future world state E_{t+1} or a trajectory $\{E_{t+1}, \ldots, E_{t+H}\}$ that satisfies or progresses toward G. Formally, the model performs:

$$\hat{E}_{t+1} = f_{\text{dec}}(f_{\text{dyn}}(f_{\text{enc}}(E_t), G))$$

Where:

- f_{enc} maps the input description into a latent state z_t.
- f_{dyn} is a latent dynamics model predicting z_{t+1} given z_t and G.
- f_{dec} optionally reconstructs a natural language description of the predicted world.

Example. Environment description: ”The robot is in a kitchen. A cup lies on the floor next to the counter.” **Goal:** ”Place the cup back on the counter.”

The system encodes this into a latent world state, simulates plausible intermediate actions (e.g., pick up, move, place), and predicts a future state:

Predicted future: ”The robot picks up the cup and places it on the counter.”

2.2 System Architecture

Our framework consists of three modular components:

1. **Semantic Translator.** A large language model encodes the current environment and goal into a latent representation z_t. z_t is a structured latent vector whose components represent key aspects of the environment, such as object properties, agent location, action affordances, or goal states. This latent space is designed to mirror the abstract semantic structure implicitly described by natural language, enabling interpretable and targeted rollout simulations. While the input and output remain in natural language, the internal simulation occurs entirely in structured or optimized latent spaces—not in token space—allowing for efficient and differentiable reasoning without textual generation at each step. The structured representation preserves semantic information extracted from language without requiring the dynamics model to relearn it. We optionally allow a learned projection of this structured latent into a more compact embedding space; see Section 2.3.
2. **Latent Dynamics Model.** A learned predictor that simulates future latent states, $z_{t+1} = f_{\text{dyn}}(z_t, G)$, conditioned on goal progress and world coherence.

3. **Verifier (Optional).** A model that scores predicted trajectories for plausibility, goal satisfaction, or alignment, enabling selection among alternatives.

This separation between interpretation (translation), simulation (dynamics), and evaluation (verification) allows for modular reasoning, controllable planning, and interpretable output generation—all within the expressive space of natural language.

2.3 Hybrid Latent Representations: From Structure to Optimization

A key benefit of our approach is that the structured latent representation z_t retains semantic structure from natural language—such as object relationships, agent state, and goal progress—without requiring the world model to relearn these mappings from scratch. By explicitly encoding these variables, we maintain semantic fidelity and interpretability.

However, structured latents may be suboptimal for predictive modeling. To support downstream learning flexibility and improve generalization, we introduce a learnable projection:

$$z'_t = f_{\text{proj}}(z_t)$$

where z'_t is a compact embedding optimized for latent trajectory prediction. The dynamics model then operates in this learned space:

$$z'_{t+1} = f_{\text{dyn}}(z'_t, a_t)$$

This projection balances symbolic structure with data-driven optimization. While z_t is aligned with semantic structure from text, z'_t can adaptively compress, reweight, or abstract these dimensions to better support accurate world modeling.

Crucially, we maintain bidirectional interpretability by decoding learned representations back into structured semantic form. A decoder $f_{\text{struct-dec}}$ maps z'_t back to $\tilde{z}_t$, a structured latent state:

$$\tilde{z}_t = f_{\text{struct-dec}}(z'_t)$$

This allows the system to retain full round-trip consistency: semantic structure from language is preserved in z_t, optimized in z'_t, and recoverable for interpretation or explanation via $\tilde{z}_t$. This cycle enables transparent inspection of learned representations while still benefiting from their flexibility during simulation.

3 Information-Theoretic Foundations

Our framework is grounded in the principle that intelligent behavior emerges from the ability to compress and predict the structure of the world. We build on Shannon's original insight—that entropy bounds the minimum number of bits needed to describe a signal—by extending it from sequences of tokens to trajectories of latent world states.

3.1 From Token Entropy to World-State Entropy

In traditional language modeling, the entropy of the next token x_{t+1} given a history $x_{1:t}$ is:

$$H(X_{t+1} \mid x_{1:t}) = -\sum_{x \in \mathcal{V}} P(x \mid x_{1:t}) \log P(x \mid x_{1:t})$$

This formulation captures the uncertainty over the next token drawn from the vocabulary $\mathcal{V}$, modeling the syntactic predictability of a token stream. However, it fails to capture the underlying dynamics of a semantically structured world.

3.2 Latent Representation of Context

We generalize this by first encoding the natural language history into a compressed latent world state. Let:

- h_t: the historical context in natural language, consisting of environment descriptions, agent observations, and goals.
- $z_t = f_{\text{enc}}(h_t)$: a latent representation summarizing the semantic world state at time t. Here, z_t is not an opaque embedding, but a structured vector space where each dimension or subspace corresponds to a semantic variable or relation in the environment (e.g., agent location, object state, goal conditions). This allows us to interpret world modeling as compression over evolving semantic configurations.

Then, the entropy over future world states becomes:

$$\begin{aligned} H(W_{t+1} \mid z_t) = & -\sum_{w \in \mathcal{W}} P(W_{t+1} = w \mid z_t) \\ & \cdot \log P(W_{t+1} = w \mid z_t) \end{aligned}$$

Here, $\mathcal{W}$ denotes the space of possible next world states, such as configurations of entities, agent actions, or environmental changes. This formulation measures the uncertainty over what happens next in the world, conditioned on its compressed semantic representation z_t.

3.3 Reasoning as Predictive Compression

Under this view, reasoning becomes the process of minimizing $H(W_{t+1} \mid z_t)$—i.e., reducing uncertainty about the future world by improving the model's predictive ability over latent trajectories. This repositions reasoning as a form of information compression: the better an agent can compress the set of possible futures, the more ”intelligent” its behavior appears.

In this framework, world models become compression engines over semantic state transitions. The latent dynamics model f_{dyn} acts as a decoder of future states, and accurate trajectory modeling becomes synonymous with understanding, simulation, and planning.

4 Planning in Latent Space

Our framework enables goal-directed reasoning by simulating future world states directly in latent space. Rather than decoding each intermediate prediction back into natural language, we operate entirely within the compressed semantic representation space, allowing for efficient rollout, evaluation, and selection of action sequences. This marks a conceptual departure from traditional language models, which simulate language evolution directly in token space, and instead moves simulation into a more structured and semantically aligned latent world.

We emphasize that decoding via f_{dec} is optional and used primarily for interpretability or explanation. During planning and latent rollout, we operate entirely within the latent space z_t, allowing for efficient simulation without the ambiguity or overhead of decoding each intermediate state.

4.1 Latent Rollout Simulation

Given an encoded latent world state $z_t = f_{\text{enc}}(h_t)$, a target goal G, and a hypothetical agent action a_t at time t we simulate future trajectories using a latent dynamics model:

$$z_{t+1} = f_{\text{dyn}}(z_t, a_t)$$

By unrolling this process over a horizon H, the system generates multiple candidate trajectories:

$$\{z_{t+1}, z_{t+2}, \ldots, z_{t+H}\}$$

Each latent represents a plausible future world state under hypothetical actions. These simulations occur entirely in latent space, avoiding the computational cost and ambiguity of decoding each step into language. These states are not decoded unless needed for explanation, visualization, or human verification via f_{dec}.

4.2 Goal Evaluation and Trajectory Selection

To evaluate whether a simulated trajectory moves toward the specified goal G, we introduce a verifier function $V(z_{t+H}, G)$ that estimates goal satisfaction or task progress. This function may be implemented as:

- A learned goal-scoring model trained on labeled examples
- A similarity function (e.g., cosine similarity) between the final latent state and an embedding of the goal
- A contrastive model trained to distinguish goal-aligned futures from distractors

The planning objective becomes:

$$a^* = \arg \max_{a_{t:t+H}} V(z_{t+H}, G)$$

That is, select the trajectory (and corresponding actions) that best satisfies the goal when projected into the future latent state.

4.3 Planning Loop

The complete planning loop consists of:

1. Encode the initial context h_t into z_t
2. Sample or enumerate candidate actions a_t
3. Roll out future latent states using f_{dyn}
4. Score each trajectory with $V(z_{t+H}, G)$
5. Select the best trajectory for execution or explanation

This architecture enables interpretable, goal-conditioned planning in language-native settings—entirely without access to a symbolic planner, grounded simulator, or sensory inputs.

5 Discussion

This work presents a theoretical framework for language-native world modeling, designed to extend the predictive paradigm of language models from token streams to structured semantic world states. Our aim is not to benchmark empirical performance, but to propose an architectural blueprint that unifies compression, simulation, and planning in structured latent space, derived from and returning to natural language.

We acknowledge that no empirical experiments are provided in this initial draft. As a sole researcher, the practical constraints of engineering, data wrangling, and tuning prohibit a full implementation at this time. However, the intent of this work is to contribute a conceptual foundation—comparable in spirit to early theoretical work in model-based reinforcement learning or probabilistic programming—that others may later instantiate and empirically explore.

Despite the lack of implementation, several key design principles emerge:

- **Latent Modularity.** By separating language understanding (semantic translation), dynamics modeling, and goal evaluation, we move away from monolithic prompting and toward compositional reasoning systems.
- **Structured-to-Optimized Flexibility.** The use of structured latent representations enables interpretable grounding of language inputs, while the learned embedding space enables effective and efficient simulation. The optional decoder for round-trip conversion preserves interpretability even after compression.
- **Predictive Compression as Cognition.** Modeling reasoning as compression over world-state entropy provides a formal lens for understanding intelligence—one rooted in predictability, not just symbol manipulation or pattern matching.

While this framework is theoretical, its components build directly on well-established technologies: transformer-based language models, learned latent dynamics (as used in Dreamer and MuZero), and vector-space similarity scoring. Each component is individually feasible, suggesting that the integrated system is well within reach given appropriate engineering resources.

Future work should explore concrete instantiations of this framework—evaluating whether structured-to-learned latent conversion improves planning efficiency, whether world-state entropy correlates with task difficulty, and whether trajectory-space simulation enables more robust goal alignment than token-level prompting alone.

This paper, therefore, should be read not as an evaluation but as an invitation: to build a new class of language-native simulators that reason in latent space, plan in structured futures, and compress the unfolding world as humans do—not token by token, but thought by thought.

References

[1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 1877–1901, 2020.

[2] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations (ICLR)*, 2020.

[3] Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97, pages 2555–2565. PMLR, 2019.

[4] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*, pages 1–22. ACM, 2023.

[5] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.

[6] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*, 2023.

[7] Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon

Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2020.

[8] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

[9] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 24824–24837, 2022.