

Decoding Future Semantics from Language Model Hidden States: Rollout Trade-offs and Causal Steering

Jaekeol Choi

`jaekeol.choi@hufs.ac.kr`

Hankuk University of Foreign Studies

Research Article

Keywords: neural language models, hidden-state representations, semantic probing, autoregressive rollout, activation steering, future prediction

Posted Date: September 30th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10973324/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Decoding Future Semantics from Language Model Hidden States: Rollout Trade-offs and Causal Steering

Jaekeol Choi^{1*}

^{1*}Division of AI Data Convergence, Hankuk University of Foreign
Studies, Yongin-si, Republic of Korea.

Corresponding author(s). E-mail(s): jaekeol.choi@hufs.ac.kr;

Abstract

Autoregressive neural language models generate text one token at a time, yet their hidden-state representations may encode semantic information about their own future continuations before those tokens are produced. This study measures the linear accessibility of that information, compares direct semantic readout with latent decoding and autoregressive rollout, and tests whether a probe-derived sentiment direction changes subsequent generation. Across four Pythia and three Qwen models and three semantic embedding spaces, a linear ridge probe consistently recovers continuation semantics from the final prompt-token hidden state. Prompt-semantic controls show that recoverability beyond prompt-level predictability is representation-dependent: the hidden-state probe exceeds a learned prompt-to-continuation baseline under MiniLM, but not under BGE or E5. In a controlled fixed-horizon comparison on Pythia-1.4B, direct semantic probing outperforms the evaluated small-data affine latent decoders and has a higher mean score than a three-token rollout, although their paired difference is not statistically distinguishable from zero; longer rollouts are stronger. Across horizons from 16 to 256 tokens, the probe becomes increasingly competitive, although the longest tested rollout remains stronger. The tested sentiment direction also shifts generation relative to matched random perturbations, while passive decodability and steering profiles do not always align across layers. These results characterize future-semantic accessibility in neural language-model representations as a horizon–compute–representation trade-off whose interpretation depends on prompt-level semantic predictability.

Keywords: neural language models, hidden-state representations, semantic probing, autoregressive rollout, activation steering, future prediction

1 Introduction

Autoregressive neural language models generate text one token at a time, yet the hidden-state computation performed before the first continuation token may already represent information about what the model itself will generate much later. Understanding how this information is organized across neural representations is relevant both to representation analysis and to inference-time control. Throughout this study, a “future continuation” refers to the continuation generated by the model itself from the current prompt, rather than to an externally supplied or human-authored continuation. If a current hidden state already exposes the semantics of this future continuation, it may provide a form of semantic lookahead without explicitly generating the intervening tokens. At the same time, demonstrating such information is decodable does not establish how much of it exceeds what is predictable from the prompt itself, how it compares with latent future decoding or autoregressive rollout, or whether a decoded direction can causally shift subsequent generation.

Previous work provides growing evidence that future behavior is partially represented before it becomes explicit. Logit- and tuned-lens methods decode latent token predictions from intermediate states, while Future Lens extends this analysis to subsequent tokens beyond the immediate next-token prediction [1, 2]. Other studies investigate whether language models prepare information needed for future computation [3], whether token activations encode upcoming paragraph content [4], and whether prompt representations reveal global properties of a future response [5]. More recent work further explores paragraph- and document-scale future information using richer activation decoders [6]. Together, these results suggest that language models can carry information predictive of their future outputs well before those outputs are fully generated.

However, three questions remain insufficiently separated. First, much of the existing work focuses on future token identities, predefined response attributes, or expressive decoders rather than asking whether the continuous semantics of a model’s extended future continuation are accessible through a simple linear readout. Such recoverability must also be distinguished from continuation semantics that are already predictable from the prompt itself. Second, probe performance is often evaluated against static or token-level controls. Such comparisons can overstate the practical strength of a probe if explicitly generating a short prefix of the same future trajectory already reveals comparable or greater semantic information. Third, passive decodability does not necessarily imply causal usefulness. Probing research has repeatedly cautioned that information recoverable from a representation need not be functionally used by the model [7, 8], while activation-steering studies show that appropriately chosen hidden-state directions can causally alter generation [9, 10]. Whether the layer that best exposes future-continuation semantics is also an effective intervention point therefore remains an empirical question.

This study addresses these issues by examining the semantics of a model’s own future continuation from the neural hidden state available before continuation generation. A linear ridge probe maps the final prompt-token representation to a sentence-level embedding of that continuation, providing a deliberately simple measure of semantic accessibility. The probe is compared with representation-level and

prompt-semantic controls, small-data affine latent-decoding configurations inspired by Tuned Lens and Future Lens, and behavioral rollout that explicitly generates the next m tokens of the same trajectory. By varying the continuation horizon H , rollout budget m , and semantic encoder E , the analysis characterizes semantic lookahead as a joint function of prediction distance, additional autoregressive computation, and semantic representation. Finally, a probe-derived sentiment direction is used for activation steering across layers to compare passive semantic accessibility with the layer-wise effect of intervention.

The experiments cover four Pythia scales and three Qwen models, together with three independently trained sentence-embedding spaces. The semantics of the models’ own future continuations are consistently linearly recoverable across models and semantic target spaces. Prompt-semantic controls qualify this result: the hidden-state probe shows a statistically reliable advantage over a learned prompt-to-continuation ridge baseline under MiniLM, while the corresponding advantages under BGE and E5 are small and not statistically distinguishable from zero. Direct paired-bootstrap contrasts further confirm that the hidden-over-prompt advantage is larger under MiniLM than under BGE or E5. In a controlled fixed-horizon comparison, direct semantic probing also outperforms the evaluated small-data affine latent-decoding configurations and has a higher mean score than a three-token behavioral rollout, although their paired difference is not statistically distinguishable from zero; longer rollouts are stronger. Across horizons from 16 to 256 tokens, the probe becomes increasingly competitive as the target extends farther into the model’s future trajectory, although the longest tested rollout remains stronger.

Causal experiments further show that the tested probe-derived sentiment direction shifts subsequent generation relative to matched random perturbations, but passive decodability and sentiment-steering profiles are not universally aligned. Some models exhibit substantial separation between the best passive readout and best-performing tested steering layers, whereas others show approximate alignment. Exploratory layer-wise results also reveal oppositely signed steering margins at selected depths. These findings support a distinction between semantic information accessible to a linear observer and the layer-wise causal efficacy of the tested sentiment intervention.

The main contributions of this study are as follows:

- It demonstrates that the semantics of a model’s own future continuation are linearly recoverable from the current hidden state across multiple model families, scales, and semantic embedding spaces. Relative to a learned prompt-semantic control, the additional hidden-state advantage is strongest and statistically reliable under MiniLM, while remaining small under BGE and E5.
- It compares continuation-level semantic probing with controlled small-data affine latent-decoding baselines inspired by Tuned Lens and Future Lens, as well as with explicit autoregressive rollout. Under the evaluated latent-decoding configuration, direct semantic readout is stronger, while its competitiveness with rollout follows a horizon–compute–representation trade-off.
- It compares passive semantic decodability with the layer-wise steerability of a probe-derived sentiment direction across models, showing that the best-performing tested readout and intervention layers may align or dissociate.

- It highlights methodological requirements for interpreting steering results, including replicated evaluation, layer-specific intervention-scale selection, and separation of passive readout quality from causal intervention efficacy.

Taken together, these results indicate that neural hidden states contain linearly accessible semantic structure predictive of a model’s own future continuation before that continuation is explicitly generated. This should not be interpreted as evidence that a complete future response is precomputed, nor as a claim about predicting an externally defined or human-authored future. Rather, semantic lookahead characterizes measurable structure related to the model’s own subsequent behavior, whose accessibility depends on prompt-level predictability, semantic representation, available autoregressive computation, and model depth.

2 Related Work

This study lies at the intersection of three lines of research: probing hidden representations, decoding information about future model behavior, and causally manipulating internal representations. The following discussion first reviews work on linear probing and its interpretational limitations, then considers methods for extracting future predictions or plans from intermediate states, and finally relates the causal steering analysis to activation-based intervention methods.

2.1 Probing and Linear Readout of Hidden Representations

Probing is widely used to characterize what information is accessible from intermediate neural representations. Early work applied simple classifiers to hidden states to recover linguistic and structural properties [11–13], while subsequent studies examined how such information emerges and changes across the layers of contextualized language models [14–16]. These studies established linear probing as a lightweight diagnostic for identifying information that can be extracted from a representation without requiring a highly expressive decoder.

The interpretation of probe performance, however, requires care. A strong probe does not necessarily imply that the model itself relies on the decoded information, and an overly expressive probe may learn structure that is only weakly reflected in the representation. Hewitt and Liang [7] therefore emphasize probe complexity and control tasks as important safeguards against misleading conclusions. Related information-theoretic approaches distinguish the amount of information recoverable from a representation from the complexity required to extract it [17, 18]. Belinkov [19] surveys these methodological issues and highlights the importance of interpreting probing primarily as evidence of accessibility rather than direct evidence of model computation.

This distinction becomes particularly important when probing semantic information. Recent work suggests that factual, relational, and other high-level properties can often be represented in approximately linear form in language-model activations [20, 21]. Such findings motivate the use of simple linear readouts when the objective is to test whether a semantic property is already directly accessible from a

hidden state, rather than merely recoverable by a powerful nonlinear decoder. At the same time, Amnesic Probing demonstrates that even clearly decodable information need not be necessary for the model’s observed behavior [8], reinforcing the distinction between representational accessibility and functional relevance.

Following this line of work, the present study uses a linear ridge probe to test whether the semantics of a model’s own future continuation are directly accessible from its current hidden state. The probe is treated as a diagnostic of semantic accessibility rather than as evidence that the model explicitly stores or uses the predicted representation, while causal intervention is evaluated separately.

2.2 Decoding Future Predictions and Plans

A growing body of work investigates whether intermediate language-model representations contain information about tokens, decisions, or higher-level properties that become explicit only later in generation. The Logit Lens projects intermediate hidden states directly into vocabulary space, while the Tuned Lens learns layer-specific affine transformations to better align intermediate representations with the model’s final prediction space [1, 22]. Related work further shows that intermediate states can sometimes be transformed into representations associated with later computation [23].

Future Lens provides a direct precedent for decoding subsequent tokens from a single hidden state and examines whether such future-related directions can also influence later generation [2]. Wu et al. [3] further distinguish between future information that is pre-computed in advance and intermediate “breadcrumbs” that facilitate later computation.

More recent studies extend this question beyond individual token identities. Pochinkov et al. [4] examine whether activations contain information about subsequent paragraph content, while Dong et al. [5] show that prompt representations can reveal global properties of upcoming responses. ParaScopes studies paragraph- and document-scale future information using more expressive decoders [6], and Patchscopes interprets hidden representations through activation patching and natural-language behavior [24]. Related evidence of semantically meaningful intermediate states has also been reported in program-generating models [25].

These studies establish that model activations can contain future-related information. The contribution of this study is therefore not this general observation, but a complementary focus on continuation-level linear semantic readout, its comparison with explicit rollout across horizons and semantic spaces, and causal validation of the corresponding directions.

2.3 Activation Steering and Causal Interpretation

A complementary line of work moves beyond passive decoding and asks whether internal representations can be manipulated to causally influence model behavior. Early studies showed that adding learned or extracted directions to hidden activations can steer generation toward desired attributes without updating model parameters [9, 26].

Representation Engineering generalizes this perspective by treating high-level concepts as directions or subspaces that can be identified and manipulated directly in activation space [27].

Subsequent methods have demonstrated such interventions across a range of behaviors. Inference-Time Intervention modifies directions associated with truthfulness [28], Contrastive Activation Addition constructs steering vectors from contrasting examples [10], and Function Vectors identify compact internal directions associated with task-level behavior [29]. Activation-based interventions have also been applied to refusal, instruction following, and knowledge-related behavior [30–32].

These approaches are closely related to, but distinct from, probing. A successful probe establishes that information is recoverable from a representation, whereas causal relevance requires that intervening on the representation changes downstream behavior. Amnesic probing demonstrates that decodable information need not be necessary for model behavior [8], while the linear representation framework of Park et al. [21] connects linear readout and intervention without requiring their strongest directions or layers to coincide.

The present study applies this distinction to future-continuation semantics, using probe-derived semantic directions to examine whether passive decodability and causal steerability exhibit the same layer-wise structure.

3 Method

We study whether the hidden state available before generation already contains information about the semantics of the model’s own future continuation, how that information compares with explicit autoregressive rollout, and whether the same representation can be used for causal intervention. Our framework therefore separates three questions: semantic recoverability, behavioral predictability under additional decoding, and causal steerability across layers.

3.1 Problem Setup

Let x denote an input prompt and let a causal language model generate a continuation

$$y_{1:H} = (y_1, y_2, \dots, y_H),$$

where H denotes the continuation horizon in tokens. Before generating y_1 , the model produces a hidden representation at the final prompt token. We denote the hidden state at layer ℓ by

$$h_\ell(x) \in \mathbb{R}^{d_\ell}.$$

Here, H denotes the realized continuation length in the primary variable-length experiments and a prespecified token horizon in the fixed-horizon analyses.

Our goal is to test whether $h_\ell(x)$ contains linearly accessible information about the semantics of the future continuation $y_{1:H}$. To operationalize continuation semantics, we use a sentence encoder E and define the semantic target as

$$z_{H,E} = E(y_{1:H}) \in \mathbb{R}^{d_E}.$$

We then learn a linear mapping from the current hidden state to the semantic representation of the future continuation:

$$\hat{z}_{\ell,H,E} = W_{\ell,H,E}h_{\ell}(x) + b_{\ell,H,E},$$

where $W_{\ell,H,E}$ and $b_{\ell,H,E}$ are estimated using ridge regression.

Semantic recoverability is measured by the cosine similarity between the predicted and true continuation embeddings:

$$S_{\text{probe}}(\ell, H, E) = \cos(\hat{z}_{\ell,H,E}, z_{H,E}).$$

A high value of S_{probe} indicates that the semantics of the model’s future continuation can be decoded from its current hidden state. Importantly, this does not by itself imply that the probe predicts future behavior better than the model’s own autoregressive generation, nor that the decoded direction is causally effective. We evaluate these questions separately using behavioral rollout baselines and hidden-state steering experiments in the following subsections.

3.2 Semantic Probe

For each prompt, we extract the hidden state at the final prompt token and use it to predict the semantic representation of the model’s future continuation. For a given layer ℓ , horizon H , and semantic encoder E , we fit a linear ridge-regression probe:

$$\min_{W,b} \sum_{i=1}^N \left\| Wh_{\ell}(x_i) + b - z_{H,E}^{(i)} \right\|_2^2 + \lambda \|W\|_F^2,$$

where $z_{H,E}^{(i)}$ denotes the semantic embedding of the continuation generated from prompt x_i , and λ is selected using held-out validation data.

Probe performance is measured on unseen prompts using the mean cosine similarity between predicted and true continuation embeddings:

$$S_{\text{probe}}(\ell, H, E) = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} \cos(\hat{z}_{\ell,H,E}^{(i)}, z_{H,E}^{(i)}).$$

We evaluate probes at multiple transformer layers because the accessibility of future semantic information may vary across model depth. Each semantic encoder defines a distinct target space, and a separate probe is trained for each encoder rather than assuming that semantic geometry is shared across embedding models.

To distinguish hidden-state recoverability from continuation predictability based on prompt semantics alone, we additionally construct prompt-semantic controls. For each encoder E , the full prompt is encoded with the same sentence encoder used to define the continuation target. The Prompt Embedding control directly compares $E(x)$ with $E(y)$, while Prompt Ridge learns a linear mapping from $E(x)$ to $E(y)$.

using the same train-validation-test split, ridge regularization grid, and validation-based selection procedure as the hidden-state probe. Prompt Ridge therefore operates entirely within the target semantic space and is intended to quantify continuation predictability from prompt semantics rather than to match the hidden-state probe in input dimensionality or parameter count.

The probe is intentionally linear. The objective is not to maximize predictive performance with a flexible decoder, but to test whether future-continuation semantics are directly accessible through a simple linear readout of the current hidden representation.

3.3 Horizon-Rollout Comparison

To assess whether semantic information decoded from the current hidden state remains competitive with explicit autoregressive generation, we compare the probe against a behavioral rollout of the same model. Let m denote the number of future tokens revealed by rollout, with $m < H$. The rollout representation is obtained by embedding the generated prefix $y_{1:m}$ in the same semantic space used for the full continuation target $y_{1:H}$.

For semantic encoder E , we define the rollout score as

$$S_{\text{rollout}}(H, m, E) = \cos(E(y_{1:m}), E(y_{1:H})).$$

We compare this score directly with the probe prediction through

$$\Delta(H, m, E) = S_{\text{probe}}(H, E) - S_{\text{rollout}}(H, m, E).$$

A positive value of $\Delta(H, m, E)$ indicates that the current hidden state provides a semantic prediction closer to the full continuation than an explicit rollout of m tokens, whereas a negative value indicates that the rollout is more informative.

This formulation separates three factors that jointly determine the comparison: the prediction horizon H , the rollout budget m , and the semantic encoder E . Increasing H asks the probe and rollout to account for more distant future content, while increasing m gives the rollout additional sequential computation and direct access to a larger portion of the generated continuation. The semantic encoder determines how partial and complete continuations are represented and compared.

We therefore interpret probe-rollout performance as a horizon-compute trade-off rather than as a binary question of whether one method universally outperforms the other. In particular, the probe requires no additional autoregressive decoding once the hidden state has been obtained, whereas rollout requires m sequential generation steps. The specific horizons, rollout budgets, and semantic encoders used in our experiments are described in Section 4.

3.4 Causal Semantic Steering

Linear recoverability alone does not establish whether a probe-derived semantic direction can causally shift subsequent generation. We therefore construct a sentiment

axis from the difference between the mean L2-normalized MiniLM embeddings of five positive and five negative anchor sentences:

$$\mathbf{u} = \frac{\bar{\mathbf{e}}_+ - \bar{\mathbf{e}}_-}{\|\bar{\mathbf{e}}_+ - \bar{\mathbf{e}}_-\|_2}.$$

For layer ℓ , this axis is mapped into hidden-state space through the corresponding refit ridge-probe weights,

$$\mathbf{d}_\ell = \frac{W_\ell^\top \mathbf{u}}{\|W_\ell^\top \mathbf{u}\|_2}.$$

The intervention is applied once during prompt prefill at the final prompt-token position,

$$\mathbf{h}'_\ell = \mathbf{h}_\ell + \alpha_\ell \mathbf{d}_\ell,$$

with no intervention during subsequent decoding. The layer-relative magnitude α_ℓ is selected using held-out pilot data as described in Section 4.4.

Generated continuations are scored against the same positive–negative semantic contrast. The reported steering margin is

$$M_\ell = [s(y_\ell^{\text{sem}}) - s(y^{\text{base}})] - [s(y_\ell^{\text{rand}}) - s(y^{\text{base}})],$$

where the matched random intervention uses a unit-norm random direction with the same α_ℓ . A positive M_ℓ therefore indicates a stronger shift in the intended sentiment direction than under an equally sized random perturbation. Exact anchor sentences, the definition of $s(\cdot)$, and full intervention details are provided in Supplementary Information (Online Resource 1), Section C.1.

4 Experimental Setup

We evaluate semantic recoverability, rollout competitiveness, and causal steerability under a shared experimental pipeline. Experiments use the same underlying 600-prompt corpus whenever applicable, while the primary recoverability analyses and the fixed-horizon analyses use different filtering and splitting protocols, as described below.

4.1 Models and Data

The experiments use base, non-instruction-tuned causal language models from the Pythia and Qwen families. The evaluated models are Pythia-410M, Pythia-1.0B, Pythia-1.4B, and Pythia-2.8B; Qwen2.5-1.5B; and Qwen3-1.7B-Base and Qwen3-4B-Base. Pythia-1.4B serves as the primary anchor model for the baseline, horizon, semantic-encoder, latent-decoding, and steering analyses, while the remaining models are used to assess robustness across architectures and scales.

The underlying prompt corpus contains 600 inputs. Half are templated prompts: 120 are evenly distributed across six fixed-template domains—story, explanation, opinion, factual, instruction, and dialogue—and the remaining 180 belong to a slot-template category consisting of five sentence frames instantiated with different topics.

The other 300 prompts are short, random-length prefixes of sentence-initial spans sampled from the first 20,000 rows of the WikiText-103 training split. Deduplication is applied only at the exact-string level. Data partitions are formed at the individual-prompt level rather than the template-family level, so the same slot-template frame may occur across training, validation, and test splits with different topic fillers; this is discussed as a limitation. The released code provides the complete prompt-construction and preprocessing procedure.

The primary experiments use variable-length continuation targets. Each prompt is greedily decoded for at most 24 new tokens. Generation is first truncated at the first end-of-sequence (EOS) token, if present, and then at the earliest token position whose decoded text ends with “.”, “!” , or “?”. Thus, the target normally corresponds to the first generated sentence; if no sentence boundary is reached, the full EOS-truncated continuation is retained. Examples with fewer than five continuation tokens are removed, yielding 549 usable examples with lengths from 5 to 24 tokens (mean = 16.41, median = 17). These examples are partitioned into 383 training, 83 validation, and 83 test examples using random seed 42. Unless otherwise noted, this fixed 549-prompt subset, determined once using the Pythia-1.4B continuation-length criterion, is reused unchanged for the primary recoverability, retrieval, semantic-encoder robustness, cross-model, rollout-baseline, and steering analyses. The hidden representation used for probing is extracted at the final prompt token before the first continuation token is generated.

The fixed-horizon analyses use a separate continuation protocol over the same original 600-prompt corpus. Sentence-boundary truncation is not applied, and only continuations reaching at least 256 tokens are retained. This produces a common subset of 577 examples, deterministically partitioned with seed 42 into 403 training, 87 validation, and 87 test examples. The same subset and split are used for all fixed-horizon conditions and for the latent future-decoding comparison. For these analyses, each continuation is truncated to $H \in \{16, 48, 96, 192, 256\}$. Using the same 577 examples across horizons ensures that observed changes with H are not confounded by changes in sample composition.

4.2 Semantic Targets and Evaluation Metrics

Future-continuation semantics are operationalized using independently trained sentence-embedding models. The primary experiments use MiniLM (**all-MiniLM-L6-v2**), while robustness analyses additionally use BGE-base (**bge-base-en-v1.5**) and E5-base (**e5-base-v2**). For each semantic encoder E , the generated continuation is mapped to the corresponding embedding space, and a separate probe is trained and evaluated in that space.

Because these encoders differ substantially in embedding geometry and anisotropy, raw cosine values are not interpreted as directly comparable measures of semantic quality across encoders. Instead, probe performance is compared with baselines defined within the same embedding space, with emphasis on relative margins and qualitative consistency across encoders.

The primary passive metric is the mean cosine similarity between the probe prediction and the true continuation embedding on the held-out test set. For the

fixed-horizon rollout analysis, we additionally report

$$\Delta(H, m, E) = S_{\text{probe}}(H, E) - S_{\text{rollout}}(H, m, E), \quad (1)$$

where a positive value indicates that the hidden-state probe is closer to the full continuation target than an m -token behavioral rollout. Paired example-level bootstrap resampling with 10,000 replicates is used to obtain 95% confidence intervals for probe–rollout and related paired semantic-performance differences.

Semantic recovery is also evaluated as a retrieval task. For each predicted embedding, the continuation embeddings from the held-out test set form the candidate pool, and the continuation corresponding to the same prompt is treated as the target item. Recall@1, Recall@5, and mean rank are reported. This evaluation complements cosine similarity by testing whether the predicted representation identifies the prompt-specific future continuation among competing candidates.

For causal steering, performance is summarized by the direction-specific margin

$$M_\ell = S_{\text{sem},\ell} - S_{\text{rand},\ell}, \quad (2)$$

where $S_{\text{sem},\ell}$ and $S_{\text{rand},\ell}$ are the base-relative semantic and matched-random intervention effects defined in Section 3.4. Steering evaluation uses five evaluation seeds, each with 60 eligible held-out prompts and independently sampled matched random directions. The probe-derived direction, intervention magnitude, and greedy generation procedure remain fixed. We average M_ℓ within each seed and report Student- t intervals over the five seed-level means. These intervals summarize variation from prompt subsampling and random-control sampling, not probe retraining or training-split resampling. The complete interval definition and steering protocol are given in Supplementary Information (Online Resource 1), Sections C.1 and D.2.

4.3 Baselines and Controls

The semantic probe is compared with three classes of alternatives: representation-level controls, learned latent future decoders, and behavioral rollout. The first two operate on the current model state without additional autoregressive generation, whereas rollout obtains additional information through sequential decoding.

The representation-level controls include four baselines. The **mean-embedding** baseline predicts the mean training continuation embedding for every test prompt, and the **random-match** baseline pairs each prompt with an unrelated continuation embedding. The **lexical term frequency–inverse document frequency (TF–IDF)** baseline predicts continuation semantics from prompt surface features. The **direct logit lens** projects the current hidden state into token space and converts the highest-scoring token candidates into semantic predictions; multiple top- k and aggregation variants are evaluated, with the validation-selected variant used for comparison. A **shuffled-target probe** additionally trains the same ridge model after randomly permuting training continuation embeddings across prompts, testing whether recoverability depends on the correct prompt–continuation correspondence.

To compare with stronger latent future-decoding approaches, the evaluation includes **Tuned Lens** [1] and a **Future-Lens-style** decoder [2]. Tuned Lens learns a layer-specific affine translator using the next-token prediction objective and is evaluated with the same token-to-semantic aggregation used for the direct logit lens. The Future-Lens-style variant trains a separate linear decoder for each future offset and predicts future tokens directly from the original pre-generation hidden state. Predictions for different offsets are therefore obtained without consuming previously predicted tokens or performing additional autoregressive steps. Semantic reconstruction is evaluated for

$$m \in \{3, 5, 10\}.$$

A multilayer perceptron (MLP) semantic probe is also evaluated as a nonlinear diagnostic, but is not treated as a primary baseline.

The strongest behavioral baseline is explicit **autoregressive rollout**. The model generates the next m tokens, and the resulting prefix is embedded in the same semantic space as the full target continuation. The primary rollout budgets are

$$m \in \{3, 5, 10, 20\},$$

with smaller subsets used in selected cross-model comparisons. Because rollout follows the same generation trajectory as the target continuation, it directly measures the benefit of additional autoregressive computation relative to extracting information from the current state alone.

For the horizon analysis, rollout is jointly evaluated over

$$H \in \{16, 48, 96, 192, 256\}.$$

The smallest tested rollout budget that matches or exceeds probe performance is recorded as the *rollout parity budget*. This quantity is used descriptively and is not extrapolated beyond the evaluated budgets.

4.4 Layer Selection and Steering Protocol

Layer selection for passive semantic decoding is treated separately from layer selection for causal intervention. Throughout the paper, layer indices follow the Hugging Face `hidden_states` convention: index 0 denotes the embedding output, while index $\ell \geq 1$ denotes the residual stream after transformer block $\ell - 1$. Steering interventions are therefore defined only for $\ell \geq 1$.

For passive probing, candidate layers are selected in advance to span model depth. Probes are fit on the training split, and the passive readout layer is selected exclusively by validation-set cosine similarity. Test performance is then evaluated at this fixed validation-selected layer. For the primary Pythia-1.4B analysis, the candidate layers are 4, 8, 12, and 20, with layer 20 selected by validation performance. The latent future-decoding comparison in Section 5.4 also uses layer 20 while following the separate fixed-horizon protocol described in Section 4.1.

Causal steering does not assume that the passive-best layer is also the most effective intervention layer. The Pythia-1.4B steering sweep covers

$$\{1, 4, 8, 12, 16, 20, 24\},$$

where layer 1 is the output of the first transformer block and layer 24 is the output of the final block. Model-specific layer sets spanning early, middle, and late depths are used for Pythia-1.0B, Qwen2.5-1.5B, Qwen3-1.7B, and Qwen3-4B under the same indexing convention. Steering evaluation follows the primary variable-length continuation protocol in Section 4.1, with a stricter eligibility requirement of at least eight continuation tokens.

For each steering layer, ridge-probe hyperparameters are first selected using the training and validation procedure. The probe is then refit on the combined training and validation sets, and the resulting weights are used to construct the semantic steering direction defined in Section 3.4. This refitting step is completed before evaluation on the held-out steering test split.

Intervention magnitude is normalized to the activation scale of each layer. The mean hidden-state norm is estimated on the validation split, and the relative coefficient k_ℓ is selected using 40 eligible pilot prompts from the validation pool, which is disjoint from the final test split. For each layer, k_ℓ is chosen to maximize the mean pilot steering margin between the semantic intervention and its matched random-direction control. The selected value is fixed before final test evaluation, and the intervention magnitude is

$$\alpha_\ell = k_\ell \cdot \mathbb{E}_{x \in \mathcal{D}_{\text{val}}} [\|h_\ell(x)\|_2].$$

The candidate scale grids and selected coefficients are reported in Supplementary Information (Online Resource 1), Section D.1.

Each probe-derived semantic intervention is paired with a unit-norm random-direction control using the same α_ℓ . Final steering evaluation follows the five-evaluation-seed procedure defined in Section 4.2: eligible test prompts and matched random directions are resampled across evaluation seeds, while the probe-derived semantic direction, intervention magnitude, and greedy generation procedure remain fixed. This protocol separates passive semantic accessibility from layer-specific causal steerability while controlling for activation scale and nonspecific perturbation effects.

4.5 Implementation and Compute

Experiments use PyTorch, Hugging Face `transformers`, `sentence-transformers`, and `scikit-learn` on a single NVIDIA RTX PRO 6000 Blackwell GPU. Generation is greedy unless otherwise stated; the sampling-robustness condition uses temperature 0.8 and $\text{top-}p = 0.95$. Hyperparameters and intervention scales are selected without test-set access. Full software, decoder, compute, artifact-reuse, and configuration details are provided in Supplementary Information (Online Resource 1), Sections D.3–D.4. Source code is available at https://github.com/maygodwithu/decoding_future_semantic.git.

Table 1 Future-semantic recoverability on Pythia-1.4B with MiniLM targets. The upper block reports linear-probe performance across layers, and the lower block reports representation-level baselines.

Method	Layer	Cosine
Semantic probe	4	0.3363
Semantic probe	8	0.3412
Semantic probe	12	0.3670
Semantic probe	20	0.4283
Mean embedding	–	0.2310
Random match	–	0.0616
Lexical TF-IDF	–	0.3612
Best logit-lens control	–	0.1812

5 Results

We first establish whether future-continuation semantics are recoverable from the current hidden state, and then progressively test how this finding changes under alternative semantic spaces, stronger behavioral baselines, longer prediction horizons, different model families, and causal intervention. This progression separates three questions: whether future semantics are decodable, whether such decoding is competitive with explicit generation, and whether the decoded information is causally usable.

5.1 Future Semantics Are Linearly Recoverable

We begin with Pythia-1.4B using MiniLM as the semantic target encoder. Table 1 reports probe performance across transformer layers together with the representation-level baselines. Semantic recoverability increases substantially in later layers: the mean test cosine rises from 0.3363 at layer 4 to 0.4283 at layer 20, which is the best layer among those evaluated.

The layer-20 probe substantially outperforms the mean-embedding and random-match controls, indicating that its predictions contain prompt-specific information beyond the global structure of the continuation embedding space. It also exceeds the lexical TF-IDF baseline by 0.0671 cosine and the strongest tested logit-lens control by 0.2471. The latter comparison is particularly important because the logit-lens controls derive token-level predictions from the same model state; their lower performance suggests that the probe is recovering semantic information that is not captured by a simple readout of currently preferred token identities.

We further evaluate the predicted embeddings as retrieval queries over the held-out continuation set. The probe retrieves the correct continuation at rank 1 for 39.76% of test examples and within the top five for 73.49%, with a mean rank of 5.65. These results are consistent with the cosine-based evaluation: the predicted representation is not merely close to a generic continuation embedding, but often remains specific

enough to identify the continuation associated with the current prompt among the 83 held-out candidates.

Taken together, these results establish the basic recoverability finding: before any continuation token is generated, the current hidden representation contains linearly accessible information about the semantics of the model’s own future output. This result alone, however, does not establish that the effect is independent of the chosen semantic encoder or that probing is stronger than allowing the model to generate a short continuation. We examine these questions next.

5.2 Recoverability Is Robust Across Semantic Target Spaces

The previous analysis defines future semantics using MiniLM embeddings. We next test whether the recoverability result depends on this particular semantic representation by repeating the analysis with BGE-base and E5-base. For all three encoders, layer 20 remains the best-performing probe layer among those evaluated; the complete comparison is reported in Supplementary Information (Online Resource 1), Section B.1. The absolute cosine scale varies considerably across semantic encoders. In particular, BGE and E5 exhibit substantially higher similarity even for prompt-independent controls; their random-match scores are approximately 0.41 and 0.69, respectively, compared with 0.0616 for MiniLM. This behavior reflects differences in embedding-space geometry and makes direct comparison of raw cosine values across encoders inappropriate. Within each embedding space, however, the hidden-state probe outperforms the mean, lexical, random-match, and token-identity controls. Retrieval provides a complementary measure that is less dependent on absolute cosine scale: Recall@5 reaches 0.735 with MiniLM, 0.651 with BGE, and 0.614 with E5. The encoder with the highest raw cosine therefore does not yield the strongest retrieval performance, further illustrating that absolute cosine magnitude is not an encoder-independent measure of recoverability. To distinguish hidden-state recoverability from semantic information already predictable from the prompt, we introduce two prompt-semantic controls. For each semantic encoder E , the full prompt is encoded with the same encoder used to define the continuation target. The Prompt Embedding control directly compares $E(x)$ with $E(y)$, whereas Prompt Ridge learns a linear mapping from $E(x)$ to $E(y)$ using the same train-validation-test split and ridge regularization-selection procedure as the hidden-state probe. Thus, Prompt Ridge operates entirely within the target semantic space, while the hidden-state probe maps from the language-model representation space into that space. The purpose of this control is to measure continuation predictability from prompt semantics rather than to match the hidden-state probe in input dimensionality or parameter count. Table 2 reports the results. The prompt controls reveal an encoder-dependent distinction. With MiniLM, the hidden-state probe exceeds Prompt Ridge by 0.028 cosine, with a paired-bootstrap 95% confidence interval of [0.010, 0.047]. For BGE and E5, the corresponding advantages are much smaller, at 0.006 and 0.002, with intervals of [−0.001, 0.014] and [−0.0001, 0.005], respectively. Thus, the MiniLM advantage is statistically distinguishable from zero, whereas no statistically reliable hidden-over-prompt advantage is detected under BGE or E5 on the primary split. To test the encoder dependence directly, we bootstrap differences

Table 2 Comparison with prompt-semantic controls on the Pythia-1.4B primary split. Prompt Ridge maps the full-prompt embedding to the continuation embedding within each semantic target space using the same ridge-selection protocol as the hidden-state probe.

Method	MiniLM	BGE-base	E5-base
Prompt embedding	0.4121	0.6213	0.7903
Prompt Ridge	0.4001	0.6970	0.8551
Hidden-state probe	0.4283	0.7034	0.8574

between the encoder-specific hidden-over-prompt advantages using identical resampled test examples. The MiniLM advantage exceeds the BGE advantage by 0.0217, with a paired-bootstrap 95% confidence interval of $[0.0072, 0.0370]$, and exceeds the E5 advantage by 0.0259, with an interval of $[0.0092, 0.0437]$. Both intervals exclude zero, directly confirming that the additional hidden-state advantage over Prompt Ridge differs across the tested semantic spaces. As a split-sensitivity check, the prompt-control analysis is repeated over ten alternative 383/83/83 train-validation-test partitions, with ridge regularization reselected on each validation split. The mean hidden-over-prompt advantages are 0.0280 ± 0.0064 for MiniLM, 0.0047 ± 0.0033 for BGE, and 0.0018 ± 0.0010 for E5. MiniLM remains positive across all ten splits, whereas BGE and E5 each become slightly negative in one split. Because the alternative partitions substantially overlap, this analysis is treated as a descriptive sensitivity check rather than as independent replication; full split-wise results are reported in Supplementary Information (Online Resource 1), Section B.2. Finally, shuffled-target probes substantially reduce recoverability for all three encoders, confirming that probe performance depends on the input-continuation correspondence rather than only on the marginal geometry of the target embeddings. Overall, future-continuation semantics are linearly recoverable across all three target spaces, but the additional advantage of the hidden state over a learned prompt-semantic baseline differs across them. The advantage is largest and most stable under MiniLM, whereas those under BGE and E5 remain small and sensitive to the evaluated partition. Semantic target space therefore affects both the apparent magnitude of semantic recoverability and how clearly it can be separated from prompt-level semantic predictability.

5.3 Short Behavioral Rollout Provides a Stronger Baseline

Although the semantic probe consistently outperforms representation-level controls, these baselines do not allow the model to perform additional autoregressive computation. The probe is therefore compared with a stronger behavioral baseline that generates a short prefix of the model’s own continuation and embeds that prefix in the same semantic space as the target. Because this prefix consists of the first m tokens of the same greedy trajectory used to define the target continuation, rollout should be interpreted as a partial-observation behavioral baseline: it measures the value of explicitly generating part of the future trajectory rather than providing an independent forecast from the initial hidden state.

Table 3 Comparison of semantic probing with token-identity and behavioral rollout baselines under the primary variable-length continuation protocol. Rollout scores measure the similarity between an m -token generated prefix and the corresponding sentence-truncated continuation target.

Model	Probe	Best logit-lens	Rollout-3	Rollout-5
Pythia-1.4B	0.4283	0.1812	0.4757	0.6371
Qwen2.5-1.5B	0.4284	0.2652	0.4434	0.6154

Table 3 reports results for Pythia-1.4B and the cross-architecture replication on Qwen2.5-1.5B under the primary variable-length continuation protocol described in Section 4.1. This setting is distinct from the fixed- $H = 16$ experiment in Section 5.4, which uses a separate common-subset split and fixed-length target.

On Pythia-1.4B, the MiniLM-based probe achieves a cosine similarity of 0.4283, substantially above the best logit-lens control at 0.1812. Under the primary variable-length target, rollout-3 obtains a higher mean cosine of 0.4757, and rollout-5 reaches 0.6371. These values describe mean performance under this condition; no significance claim is made for the probe-rollout differences in Table 3. In the separate fixed- $H = 16$ comparison of Section 5.4, the ordering of the means differs: the probe reaches 0.450 and rollout-3 reaches 0.406, while their difference is not statistically significant.

A similar mean-level pattern is observed for Qwen2.5-1.5B. The semantic probe reaches 0.4284, clearly exceeding the strongest logit-lens baseline at 0.2652, while rollout-3 obtains 0.4434 and rollout-5 reaches 0.6154. The replication indicates that a short behavioral rollout can provide a substantially stronger reference than token-identity controls across both model families.

These results refine the interpretation of semantic recoverability. The probe demonstrates that future-continuation information is accessible from the current hidden state before generation, but this should not be interpreted as universal superiority over the model’s own near-term behavior. Rollout and probing also incur different computational costs: the probe uses an already available hidden state, whereas an m -token rollout requires m additional sequential decoding steps. The following analyses therefore compare direct semantic readout with stronger latent future decoders and examine how probe-rollout competitiveness changes with prediction horizon, rollout budget, and semantic representation.

5.4 Comparison with Existing Latent Future-Decoding Methods

The semantic probe is next compared with stronger methods that decode future information directly from the current hidden state. Figure 1 reports results at $H = 16$ on the 87 held-out test examples from the fixed-horizon split described in Section 4.1. Because this comparison follows the fixed-horizon protocol, its probe score differs slightly from the primary recoverability result in Table 1, which uses the variable-length continuation protocol and an 83-example test split.

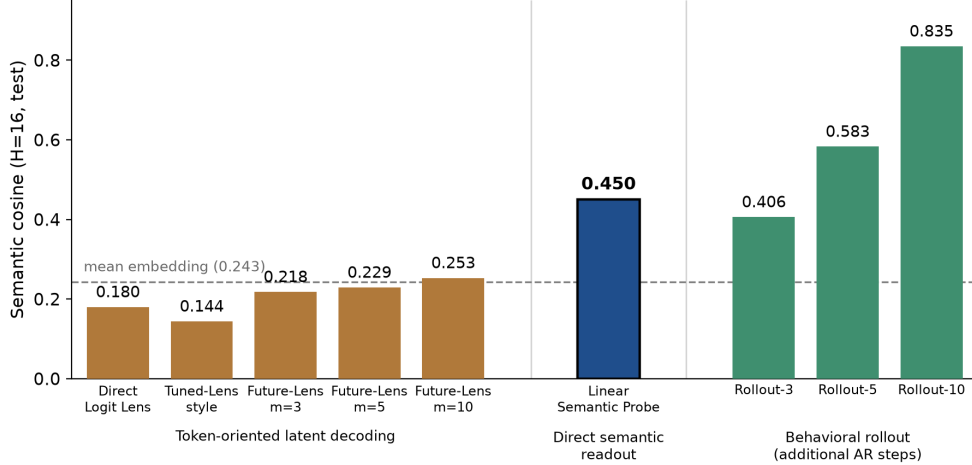


Fig. 1 Comparison of future-decoding methods at $H = 16$ on the fixed-horizon test split. Token-oriented latent decoders and the linear semantic probe require no additional autoregressive generation, while rollout explicitly generates 3, 5, or 10 additional tokens. The dashed line denotes the mean-embedding baseline.

The comparison includes the direct logit lens, a Tuned-Lens-style affine translator inspired by Tuned Lens [1], and Future-Lens-style decoders [2]. The Tuned-Lens-style baseline is trained specifically for this experiment rather than using a publicly pretrained Tuned Lens checkpoint. All latent-decoding methods operate without additional autoregressive generation, whereas behavioral rollout explicitly generates m additional tokens.

The linear semantic probe achieves a cosine similarity of 0.450, substantially exceeding the direct logit lens (0.180), the Tuned-Lens-style affine translator (0.144), and Future-Lens-style decoding with $m = 3, 5$, and 10 (0.218, 0.229, and 0.253, respectively). The probe advantage over Future-Lens-style decoding remains significant at every tested budget; at $m = 10$, the paired-bootstrap 95% confidence interval for the probe advantage is [0.164, 0.229].

The Tuned-Lens-style affine translator performs below the direct logit lens in this setting. This result should be interpreted narrowly: the translator is trained from scratch on the available 403-example training split using a next-token cross-entropy objective, whereas the reported comparison evaluates continuation-level semantic similarity. It therefore characterizes the tested affine-decoder configuration rather than establishing a general performance ordering between Tuned Lens and direct logit-lens methods.

Explicit autoregressive computation produces a different pattern. Rollout-3 reaches 0.406; although its mean score is below the semantic probe, the difference is not statistically significant. Rollout-5 and rollout-10 increase performance to 0.583 and 0.835, respectively. Thus, under this fixed- $H = 16$ setting, direct continuation-level semantic readout is substantially stronger than the tested token-oriented latent-decoding alternatives without additional autoregressive decoding steps, while sufficiently long behavioral rollout achieves higher continuation-level semantic similarity.

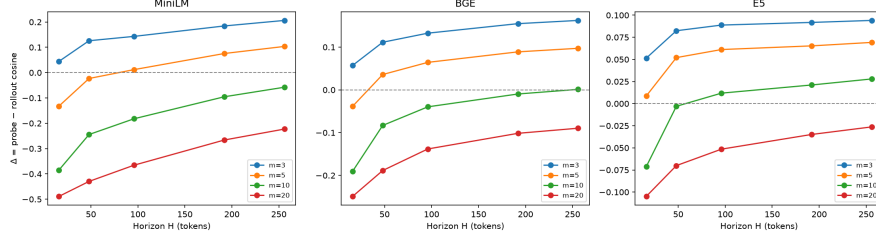


Fig. 2 Probe-rollout margin across prediction horizons, rollout budgets, and semantic encoders. Positive values indicate that the hidden-state probe is closer to the full continuation target than the corresponding behavioral rollout. The observed mean margin increases with horizon for all 12 encoder-budget combinations. All conditions use the same 87 held-out test examples.

As an additional diagnostic, a nonlinear MLP semantic probe reaches only 0.281. This result indicates that the tested nonlinear readout does not improve over the linear probe under the present configuration, but it does not imply that future semantics are intrinsically limited to a linear representation.

5.5 Horizon, Rollout Budget, and Semantic Space Jointly Determine Probe Competitiveness

The previous results show that behavioral rollout can be highly competitive with the hidden-state probe, but this comparison depends on how far into the future the semantic target extends and how much additional generation is allowed. We therefore evaluate the probe-rollout margin

$$\Delta(H, m, E) = S_{\text{probe}}(H, E) - S_{\text{rollout}}(H, m, E)$$

across prediction horizons

$$H \in \{16, 48, 96, 192, 256\},$$

rollout budgets

$$m \in \{3, 5, 10, 20\},$$

and three semantic target encoders. All conditions use the same 87 held-out examples from the common fixed-horizon split, preventing changes in sample composition from confounding the horizon comparison.

Figure 2 shows a consistent descriptive horizon effect. Across all 12 encoder-budget combinations, the observed mean probe-rollout margin increases monotonically as the target horizon becomes longer. Thus, for a fixed rollout budget, explicit generation becomes progressively less advantageous relative to the probe as the semantic target extends farther into the future. This pattern is observed for MiniLM, BGE, and E5 despite their different embedding geometries. Full paired-bootstrap confidence intervals for all conditions are reported in Supplementary Information (Online Resource 1), Section B.3.

The observed rollout parity budget is also non-decreasing with horizon for all three semantic encoders. For MiniLM, five generated tokens are sufficient at $H = 16$ and $H = 48$, whereas ten are required from $H = 96$ onward. BGE requires up to 20 tokens at $H = 256$, while E5 requires at least 10 tokens at the shortest horizon and 20 tokens from $H = 96$ onward. The complete parity table is reported in Supplementary Information (Online Resource 1), Section B.3.

The semantic encoder changes where this parity point occurs. MiniLM, BGE, and E5 differ in how embeddings of short prefixes relate to embeddings of complete continuations, shifting the rollout budget required to reach the probe. The apparent strength of behavioral rollout therefore depends not only on rollout length but also on the semantic space in which partial and full continuations are compared.

The longest tested rollout nevertheless remains stronger throughout the evaluated range. For $m = 20$, the probe-rollout margin is negative at every horizon and for all three semantic encoders, and its paired-bootstrap 95% confidence interval remains below zero in all 15 encoder-horizon conditions. Thus, no crossover against the strongest tested rollout is observed through $H = 256$; instead, its advantage narrows systematically as horizon increases.

This pattern yields a horizon-compute trade-off. The probe obtains its semantic prediction directly from a hidden state already available before continuation generation, whereas rollout requires additional sequential decoding. Moreover, the observed marginal semantic gain per additional rollout token decreases monotonically across all 15 encoder-horizon conditions. Taken together, these results do not support universal superiority of semantic probing over explicit rollout. Rather, the probe provides a computation-efficient semantic estimate whose relative competitiveness increases for more distant targets, while sufficiently long rollout remains more accurate within the tested regime.

5.6 Semantic Recoverability Generalizes Across Models and Decoding Strategies

We next examine whether future-semantic recoverability is specific to the Pythia-1.4B anchor model. Table 4 summarizes the best passive probe performance across four Pythia sizes and three Qwen models. The effect remains remarkably stable across the tested models: probe cosine ranges from 0.4020 to 0.4312 despite substantial differences in parameter count and architecture.

Within the Pythia family, all four model sizes retain a substantial advantage over their mean-embedding baselines, with margins of 0.195, 0.178, 0.197, and 0.207 for 410M, 1.0B, 1.4B, and 2.8B, respectively. The absolute probe scores show no clear monotonic relationship with model size: the 1.0B model yields the lowest value at 0.4020, while the 2.8B model is only slightly above the 410M and 1.4B models. Thus, within the tested range, semantic recoverability appears to be a broadly shared property rather than one that emerges only beyond a particular scale.

The finding also transfers across architecture families. Qwen2.5-1.5B achieves a probe cosine of 0.4284, nearly identical to the Pythia-1.4B anchor value of 0.4283, while outperforming its mean-embedding, random-match, lexical, and logit-lens baselines. Qwen3-1.7B and Qwen3-4B similarly achieve probe cosines of 0.4274 and

Table 4 Future-semantic recoverability across model families and scales under the primary variable-length continuation protocol. Probe cosine is reported at the best passive layer evaluated for each model using the MiniLM semantic target space.

Family	Model	Probe cosine
Pythia	410M	0.4260
	1.0B	0.4020
	1.4B	0.4283
	2.8B	0.4312
Qwen2.5	1.5B	0.4284
Qwen3	1.7B	0.4274
	4B	0.4094

0.4094, respectively, and both remain above their strongest tested weak and token-identity controls. These replications indicate that the effect is not specific to Pythia’s architecture or training configuration.

We further test whether the result depends on deterministic greedy decoding. On Pythia-1.4B, replacing greedy decoding with sampling at temperature 0.8 and top- $p = 0.95$ reduces the probe cosine from 0.4283 to 0.3508. However, the corresponding mean-embedding baseline also decreases from 0.2310 to 0.2042, and retrieval Recall@5 remains high at 0.699 compared with 0.735 under greedy decoding. Thus, stochastic generation weakens predictability, as expected because the continuation contains additional sampling variability, but the hidden state still carries substantial information about the semantics of the sampled future output.

Overall, future-semantic recoverability persists across all tested model sizes, two architecture families, and both deterministic and stochastic decoding. At the same time, neither model size nor architecture exhibits a simple monotonic relationship with probe performance. The next question is whether this broadly decodable information is merely correlational or can also be used to causally influence subsequent generation.

5.7 A Probe-Derived Sentiment Direction Causally Steers Future Semantics

We next test whether the probe-derived sentiment direction can be used to causally influence subsequent generation rather than serving only as a passive readout. Figure 3 summarizes the final sentiment-steering margins across all evaluated models. In every case, intervention along the probe-derived sentiment direction produces a larger mean shift in the intended direction than the matched random-direction control, providing evidence that this direction is not purely diagnostic.

Several patterns stand out. First, the mean steering margin is positive for every tested model, and every replicate-level interval remains above zero. Thus, the advantage of the probe-derived sentiment direction over matched random perturbations is

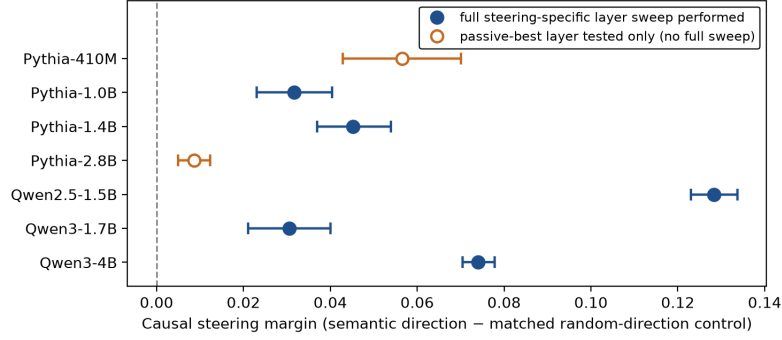


Fig. 3 Sentiment-steering margins across models. Points show mean steering margins and horizontal bars show 95% Student- t intervals over five evaluation-seed means. Filled markers denote models for which a full steering-specific layer sweep was performed; hollow markers denote models evaluated only at the passive-best tested layer. Positive values indicate that intervention along the probe-derived sentiment direction changes continuation semantics more strongly in the intended direction than a matched random-direction control.

consistent across the evaluated steering replicates and is not confined to a single architecture or scale. Second, the magnitude of the effect varies substantially across models. Qwen2.5-1.5B shows the largest observed steering margin at 0.1283, while Pythia-2.8B shows the smallest at 0.0087. Third, these differences do not follow a simple scaling trend: for example, Pythia-410M yields a larger margin than Pythia-1.0B and Pythia-2.8B, while Qwen3-4B exceeds Qwen3-1.7B. The complete numerical summary is provided in Supplementary Information (Online Resource 1), Section C.2.

These results support the conclusion that the tested probe-derived sentiment direction is not purely diagnostic. Intervening along this direction systematically biases subsequent generation in the intended sentiment direction relative to equally scaled random perturbations. At the same time, the variation in effect size across models suggests that causal steerability of this direction is not determined solely by passive semantic decodability. We therefore next examine how sentiment-steering effectiveness depends on intervention layer and how it relates to passive semantic decodability.

5.8 Passive Decodability and Sentiment Steering Differ Across Tested Layers

The preceding results show that the tested probe-derived sentiment direction can causally influence future generation, but they do not imply that the layer with the strongest passive semantic signal is also the most effective tested intervention point. We therefore compare passive probe performance and sentiment steering across model depth. Figure 4 shows the full layer profile for the Pythia-1.4B anchor model.

For Pythia-1.4B, passive decodability increases substantially toward later layers, reaching its maximum cosine of 0.4283 at layer 20. Sentiment steering remains near zero through most early and middle layers before increasing sharply in the final portion of the network. Layer 20 yields the largest observed mean steering margin of 0.0452 with a replicate-level 95% interval of [0.0367, 0.0537], while layer 24 also produces a positive mean margin of 0.0311. A direct comparison between layers 20 and 24

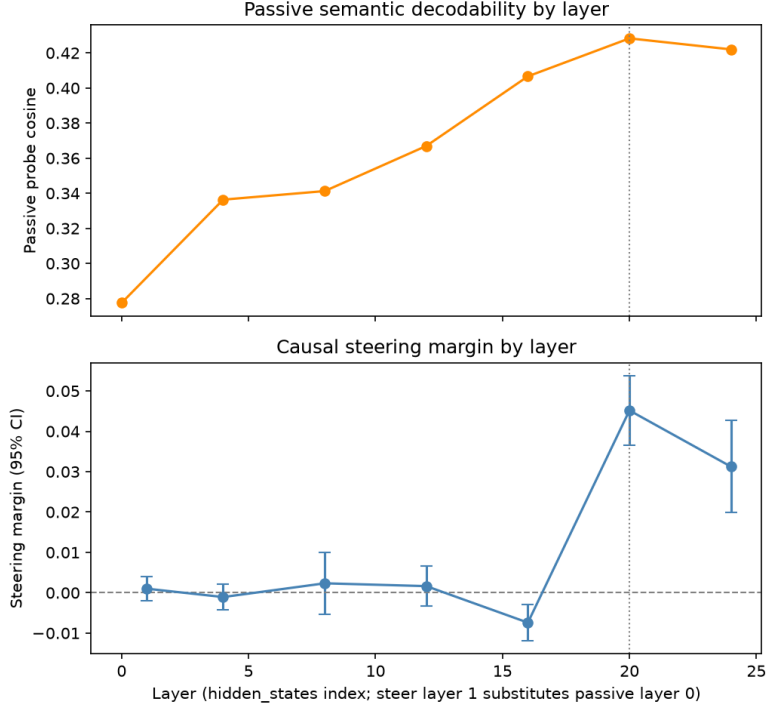


Fig. 4 Passive semantic decodability and sentiment steerability across layers in Pythia-1.4B. The upper panel shows passive probe cosine, and the lower panel shows mean sentiment-steering margins with replicate-level 95% Student- t intervals over five evaluation-seed means. The dotted line marks layer 20, which yields both the highest passive probe score and the largest observed steering margin. Despite this overall alignment, the interval for layer 16 lies below zero while passive decodability remains relatively high.

does not show clear separation under the evaluated replicates, suggesting a relatively broad region of effective late-layer intervention rather than a sharply isolated layer preference.

The correspondence is nevertheless far from monotonic. In particular, layer 16 achieves a passive probe cosine of 0.407, close to the best passive layers, but produces a negative mean steering margin of approximately -0.007 , with its replicate-level interval entirely below zero. Across the seven tested layers, the Pearson correlation between passive probe cosine and steering margin is 0.615 ($p = 0.14$), indicating a positive but statistically inconclusive association. Thus, high passive decodability alone is insufficient to predict the sign or magnitude of sentiment steering.

This relationship also differs across models. Table 5 compares models for which a steering-specific layer scan was performed. Pythia-1.0B exhibits a strong dissociation: its passive-best layer is 13, whereas the largest observed sentiment-steering margin occurs at layer 6. Qwen2.5-1.5B likewise achieves its best passive decoding at layer 24 but its largest observed steering margin at layer 28. The discrepancy is particularly large for Qwen3-1.7B, whose passive-best layer is 24 while its strongest tested

Table 5 Comparison of passive-best and best-performing tested sentiment-steering layers for models with steering-specific layer scans. “Aligned” denotes identical or closely neighboring layers, while “dissociated” denotes a substantial shift between the two best-performing tested layers.

Model	Passive-best	Largest-margin tested layer	Margin	Relationship
Pythia-1.0B	13	6	0.0317	Dissociated
Pythia-1.4B	20	20	0.0452	Aligned
Qwen2.5-1.5B	24	28	0.1283	Dissociated
Qwen3-1.7B	24	14	0.0306	Dissociated
Qwen3-4B	32	31	0.0740	Approximately aligned

sentiment-steering effect occurs at layer 14. In contrast, Pythia-1.4B and Qwen3-4B show approximate alignment between the two layer profiles.

The layer dependence is further complicated by exploratory sign reversals. In addition to the negative margin at layer 16 of Pythia-1.4B, Pythia-1.0B exhibits a negative steering margin at layer 4 whose replicate-level interval also lies below zero. Because multiple layers are examined with independently pilot-selected intervention scales, these sign reversals are treated as exploratory observations rather than confirmatory evidence for a layer-specific mechanism. They nevertheless illustrate that strong passive semantic decodability does not guarantee that intervention along the tested sentiment direction will move generation in the expected direction at every depth.

Taken together, the results support a distinction between *where future semantics can be read out* and *where the tested sentiment direction is most effective for intervention*. These properties sometimes align, as in Pythia-1.4B and Qwen3-4B, but can diverge substantially in other models. Passive semantic decodability and sentiment steerability should therefore be treated as related but separate, model- and layer-dependent properties rather than assuming that the best probing layer is automatically the best intervention layer.

6 Discussion

The results reveal that future-continuation semantics are accessible from current hidden states, but the interpretation of this accessibility depends critically on how it is measured and how the representation is used. In this section, we discuss what semantic recoverability does and does not imply, how explicit rollout changes the notion of predictive advantage, and why passive decoding and causal intervention should be treated as related but distinct questions.

6.1 What Does Future-Semantic Recoverability Mean?

The central passive result is that the model’s current hidden state contains linearly accessible information about the semantics of its own future continuation. This pattern is observed across multiple model families and scales and remains visible under MiniLM, BGE, and E5 target spaces. Shuffled-target controls further show that recovery depends on the correspondence between a prompt representation and its own continuation rather than only on the marginal geometry of continuation embeddings.

The stronger prompt-semantic controls, however, qualify how this result should be interpreted. Under MiniLM, the hidden-state probe significantly outperforms a ridge model that predicts continuation semantics directly from the prompt embedding, indicating continuation-predictive structure beyond that captured by an external semantic representation of the prompt. Under BGE and E5, the corresponding differences are small and not statistically significant. Thus, continuation semantics are recoverable across embedding spaces, but the extent to which this recoverability exceeds prompt-level semantic predictability is representation-dependent. This also reinforces that raw cosine values should not be treated as an encoder-independent measure of future information.

The latent-decoding comparison provides a complementary perspective. Under the evaluated Pythia-1.4B, $H = 16$, MiniLM condition, direct semantic probing outperforms the tested Tuned-Lens-style affine translator and Future-Lens-style semantic reconstruction procedures. This result is specific to the evaluated small-data affine configurations and does not establish a general ordering over latent-decoding methods. Likewise, the failure of the tested nonlinear MLP probe to improve over ridge is consistent with simple linear accessibility in this setting, but does not imply that the underlying representation is intrinsically linear.

These findings should therefore not be interpreted as evidence that the model stores a fully specified future sequence before generation begins. The probe predicts a semantic representation rather than an exact token trajectory, and multiple continuations can share similar semantics. The most robust conclusion is that future-continuation semantics are partially accessible from the pre-generation hidden state, while how much of that accessibility is distinct from prompt semantics depends on the representation used to define meaning. This leaves open both the internal mechanism underlying the readout and how directly the recovered semantic structure determines subsequent token-level generation.

6.2 Semantic Lookahead as a Horizon–Compute Trade-off

The rollout experiments substantially refine how semantic lookahead should be interpreted. Against weak representation-level controls and stronger latent future-decoding methods, the direct semantic probe is clearly competitive. In particular, the tested Tuned-Lens-style affine translator and Future-Lens-style decoding configurations remain well below the probe without additional autoregressive generation. However, once the comparison allows the model to generate additional tokens, behavioral rollout becomes progressively stronger. At $H = 16$, the probe is statistically comparable to a three-token rollout, whereas five- and ten-token rollouts clearly outperform it. Thus, semantic probing should not be framed as universally more accurate than explicit generation.

Under the evaluated conditions, the observed ordering places the tested latent-decoding configurations below direct semantic readout, while short rollout is competitive with the probe and longer rollout is stronger. The first two approaches extract information from the current hidden state without additional autoregressive steps, whereas rollout obtains new information through sequential computation. Directly decoding continuation-level semantics therefore appears substantially more effective

than first reconstructing future token identities and then converting them into a semantic representation, at least under the evaluated conditions.

The more informative comparison is the trade-off between prediction horizon and additional autoregressive computation. As shown in Figure 2, for every tested semantic encoder and rollout budget, the probe-rollout margin increases as the target horizon grows. A fixed number of revealed tokens therefore becomes progressively less informative about increasingly distant continuation semantics. Consistent with this trend, the smallest rollout budget required to match the probe is non-decreasing with horizon across all three semantic spaces.

A full crossover against the strongest tested rollout is not observed. An $m = 20$ rollout remains more similar to the full continuation than the probe for every evaluated horizon up to $H = 256$. The result therefore does not support the stronger claim that hidden-state probing eventually dominates sufficiently long rollout. Instead, it shows that the relative advantage of explicit generation narrows systematically as the semantic target extends farther into the future.

This distinction matters computationally because the approaches obtain information in different ways. Once the current hidden state has been computed, latent decoding and semantic probing require no additional autoregressive generation, whereas rollout requires m sequential decoding steps. At the same time, the marginal semantic gain from additional rollout tokens decreases across all tested encoder-horizon combinations. Longer rollout remains effective, but increasingly more sequential computation is required to maintain its advantage for distant targets.

The semantic target space introduces an additional dimension to this trade-off. MiniLM, BGE, and E5 differ in how short prefixes relate to complete continuations in embedding space, shifting the rollout budget required to reach parity with the probe. Consequently, the apparent strength of rollout is not a property of the language model alone; it also depends on the semantic representation used for evaluation.

Taken together, these findings characterize semantic lookahead as a horizon-compute-representation trade-off. Direct semantic probing provides a strong zero-generation estimate of future continuation semantics, while explicit rollout exchanges additional sequential computation for increasingly direct evidence about the future. The relevant question is therefore not which method is universally superior, but how much additional computation is required to obtain a desired level of semantic information at a given prediction horizon.

6.3 Passive Readout and Causal Intervention Are Distinct

The causal steering results show that the tested probe-derived sentiment direction is not purely diagnostic. Across all evaluated models, intervention along this direction produces a positive mean steering margin relative to a matched random-direction control. This indicates that the tested sentiment direction can causally shift subsequent generation relative to matched random perturbations.

However, the layer-wise analyses show that passive semantic decodability and the effectiveness of the tested sentiment intervention should not be treated as the same property. Under the evaluated pilot-selected intervention scales, the layer with the strongest passive probe performance is not always the layer with the largest observed

sentiment-steering margin. Pythia-1.0B, Qwen2.5-1.5B, and Qwen3-1.7B show clear shifts between their passive-best and best-performing tested steering layers, whereas Pythia-1.4B and Qwen3-4B show approximate alignment. The observed relationship therefore varies across models and tested layers rather than following a universal pattern.

The Pythia-1.4B layer profile in Figure 4 further illustrates why a high passive probe score should not be interpreted as direct evidence of effective sentiment steering. Layer 16 retains relatively strong passive recoverability, yet the corresponding steering margin is negative, with its replicate-level interval lying below zero. A similar oppositely signed margin appears at layer 4 of Pythia-1.0B. Because these observations arise from layer-wise analyses with independently pilot-selected intervention scales, they are treated as exploratory rather than as evidence for a specific layer-wise causal mechanism. They nevertheless illustrate that a representation can support semantic readout while intervention along the corresponding probe-derived sentiment direction fails to move generation in the expected direction.

One possible interpretation is that linear decodability identifies directions correlated with future semantics without guaranteeing that those directions coincide with the local causal geometry governing generation. Information may be readable from a representation because it is redundantly encoded, transformed for downstream use, or correlated with other variables that support prediction. Intervention, by contrast, requires that perturbation along a particular direction propagate through subsequent computation in a way that affects output semantics. Our experiments do not distinguish among these mechanisms, but they show empirically that readout quality alone is insufficient to determine intervention efficacy.

This distinction also cautions against selecting steering layers solely from passive probe performance. For Pythia-1.0B, evaluation at the passive-best layer produced a near-zero steering margin, whereas a steering-specific sweep identified a substantially larger observed margin at an earlier layer. In contrast, Pythia-1.4B shows approximate alignment between the passive-best layer and the best-performing tested steering layer under the evaluated pilot-selected intervention scales. Intervention-layer selection should therefore be treated as a separate empirical problem rather than inferred directly from passive probe quality.

Overall, the results show that passive semantic accessibility and the effectiveness of the tested sentiment intervention need not exhibit the same layer-wise pattern. Their degree of alignment varies across models and tested layers, with exploratory analyses also revealing oppositely signed margins at selected depths. These observations motivate separate intervention validation when passive probe performance is used to select steering layers.

6.4 Methodological Lessons from Steering Evaluation

The steering experiments also provide a methodological lesson about how causal-intervention results can be distorted by intervention scale, limited evaluation replication, or the assumption that a single layer is representative of the whole model. In the present experiments, each of these factors materially affected the interpretation of initially anomalous results.

The first case involved Pythia-2.8B. Under an initial single evaluation with a fixed raw intervention magnitude, its steering margin was only 4.23×10^{-5} , suggesting little measurable steerability. Norm-relative scaling increased the margin but still left the effect small. Under the finalized protocol, however, Pythia-2.8B achieves a mean steering margin of 0.0087 with a replicate-level 95% interval of [0.0050, 0.0124]. Because these intervals summarize variation across evaluation seeds that resample both eligible prompts and matched random controls, this result is better interpreted as showing that the initial near-zero observation was not representative of the finalized evaluation procedure, rather than as evidence of a scale-specific steering failure.

Pythia-1.0B revealed a different issue. At the passive-best layer, the mean steering margin remained approximately zero even after norm-relative scale selection,

$$-0.00002, \quad \text{replicate-level 95\% interval} = [-0.0031, 0.0030].$$

A steering-specific layer sweep, however, identified a much larger observed margin at layer 6: 0.0317 with a replicate-level 95% interval of [0.0230, 0.0404]. The apparent model-level failure was therefore better explained by the choice of intervention layer than by an absence of measurable sentiment-steering effects. This finding motivated the broader comparison between passive semantic decodability and sentiment-steering effectiveness in Section 5.8.

These cases suggest two practical requirements for activation-steering studies. First, intervention effects should be evaluated across multiple evaluation replicates when their magnitude is small relative to run-to-run variation, with the source of that variation stated explicitly. Second, passive probe quality should not be used as the sole criterion for selecting an intervention layer. A layer can support strong passive readout yet weak or oppositely signed intervention effects, while another layer may yield a substantially larger steering margin.

More broadly, anomalous steering results should be diagnosed before being interpreted as architectural or scaling effects. Here, one apparent failure changed after scale normalization and replicated evaluation, while another changed after layer-specific intervention analysis. The resulting conclusion is therefore more limited but also better supported: the tested probe-derived sentiment intervention produces positive mean margins relative to matched random controls across all evaluated models, while the observed magnitude and best-performing tested layer vary substantially across models under the pilot-selected intervention scales.

7 Limitations and Future Work

Several limitations define the scope of the conclusions. First, the horizon-rollout analysis covers continuation horizons up to $H = 256$ tokens and rollout budgets up to $m = 20$. Although the probe becomes increasingly competitive as the target horizon grows, the strongest tested rollout remains superior throughout the evaluated range, so the results do not establish an eventual crossover. In addition, the fixed-horizon analysis uses the common subset of prompts whose continuations extend to at least 256 tokens before termination. This design keeps the evaluated examples identical across horizons, but conditions the analysis on relatively long generation trajectories;

the observed horizon trends may therefore not generalize to prompts whose continuations terminate earlier. Extending both horizons and rollout budgets while including shorter trajectories would clarify the generality of the observed trade-off.

Second, continuation semantics are operationalized using MiniLM, BGE-base, and E5-base. Recoverability is observed across all three spaces, but both absolute cosine values and rollout parity depend strongly on embedding geometry. More importantly, the hidden-state probe shows a statistically reliable advantage over the learned prompt-semantic control under MiniLM, but not under BGE or E5. Thus, the amount of continuation-predictive information that can be separated from prompt-level semantic predictability is itself representation-dependent. Future work should consider additional semantic representations, including task-specific or multidimensional targets.

Third, the prediction targets are the language models’ own generated continuations. This design directly measures information about each model’s future behavior, but does not establish whether the same hidden states predict human-written references, externally desired outputs, or alternative valid continuations. The prompt set is also relatively modest and combines templated and WikiText-derived inputs. Because the partitions are formed at the prompt level, the same slot-template frame can occur across training, validation, and test splits with different topic fillers. Performance may therefore overestimate generalization to previously unseen template structures. Broader tasks and more diverse prompt distributions would strengthen claims of generality.

Fourth, the evaluated language models are base, non-instruction-tuned models from the Pythia and Qwen families. Instruction-tuned and conversational models may organize future-semantic information differently because of alignment and instruction-following objectives. Extending the analysis to instruct and chat models, as well as additional architecture families, is therefore an important direction.

Fifth, the comparison with latent future-decoding methods is concentrated on the Pythia-1.4B anchor condition, with the primary comparison conducted at $H = 16$ in the MiniLM semantic space. The Future-Lens-style comparison is a continuation-level semantic adaptation rather than an exact reproduction of the original method’s native evaluation protocol, and the Tuned-Lens-style affine translator is trained specifically on the available small-data split. The MLP probe is likewise included only as a nonlinear diagnostic rather than an exhaustive nonlinear baseline. Broader cross-model, cross-horizon, and native-metric comparisons would provide a more complete characterization of these decoding approaches.

Sixth, the causal analysis is limited to a single probe-derived sentiment direction, and steering-specific layer sweeps are not equally exhaustive across models. The finalized steering evaluation uses five evaluation seeds whose variation reflects held-out prompt subsampling and matched random-control sampling while keeping the learned direction and intervention scale fixed; the resulting replicate-level intervals therefore do not capture uncertainty from probe retraining or alternative train-validation splits. Passive probing is also evaluated primarily under a fixed split. Moreover, oppositely signed steering margins at selected layers are exploratory observations obtained after examining multiple layers with independently pilot-selected intervention scales and are

not interpreted as evidence for a specific mechanism. Future work should test additional semantic directions, repeat probe and steering estimation across independent data splits, and examine how interventions propagate across layers.

Overall, these limitations constrain the scope of the interpretation rather than the observed empirical patterns. Future work should broaden model and task coverage, extend the horizon–compute regime, evaluate additional semantic representations and latent decoders, and more fully characterize the relationship between passive semantic accessibility and targeted causal intervention.

8 Conclusion

This study investigated whether autoregressive language-model hidden states contain information predictive of the semantics of their own future continuations before those continuations are explicitly generated. Across Pythia and Qwen models and multiple semantic embedding spaces, these continuation semantics are consistently recoverable with a simple linear probe. Prompt-semantic controls qualify this result: the hidden-state probe shows a statistically reliable advantage over a learned prompt-to-continuation baseline under MiniLM, whereas the corresponding advantages under BGE and E5 are small and not statistically distinguishable from zero. Direct encoder contrasts and repeated-split analyses further indicate that the additional hidden-over-prompt signal is strongest and most stable under MiniLM. Thus, semantic recoverability from the hidden state is robust, while the extent to which it exceeds prompt-level semantic predictability depends on the representation used to define the target.

Comparison with alternative decoding strategies further characterizes this accessibility. In the evaluated fixed-horizon Pythia-1.4B condition, direct semantic probing outperforms the tested small-data Tuned-Lens-style and Future-Lens-style affine decoding configurations and has a higher mean score than a three-token behavioral rollout, although their paired difference is not statistically distinguishable from zero; longer rollouts are stronger. Across horizons from 16 to 256 tokens, the probe becomes increasingly competitive relative to fixed rollout budgets, and the rollout budget required to match it is non-decreasing. Semantic accessibility of the model’s own future continuation is therefore better understood as a horizon–compute–representation trade-off than as unconditional superiority of either probing or explicit generation.

The tested probe-derived sentiment direction also shifts subsequent generation relative to matched random perturbations. Under the evaluated pilot-selected intervention scales, the best-performing tested steering layer aligns with the best passive readout layer in some models and differs substantially in others. Exploratory sign reversals at selected depths further illustrate that strong passive decodability does not guarantee an intervention effect in the expected direction. These findings establish a distinction between passive semantic accessibility and the effectiveness of the tested sentiment intervention, without implying a universal layer-wise relationship across semantic directions.

Taken together, the results support the presence of linearly accessible semantic structure predictive of a model’s own subsequent generation before that generation is explicitly produced. They do not imply that a complete future response is precomputed or that the same conclusions necessarily apply to externally supplied or human-authored continuations. Rather, semantic lookahead characterizes measurable structure related to the model’s own future behavior whose interpretation depends on prompt-level predictability, semantic representation, available autoregressive computation, and intervention setting.

Supplementary Information. Additional probe results, robustness analyses, causal-steering details, and reproducibility information are provided in Supplementary Information (Online Resource 1).

Acknowledgements. This work was supported by Hankuk University of Foreign Studies Research Fund of 2026.

Statements and Declarations

Funding. This work was supported by Hankuk University of Foreign Studies Research Fund of 2026.

Competing interests. The author has no relevant financial or non-financial interests to disclose.

Ethics approval and consent to participate. Not applicable. This study involved computational experiments using publicly available text data and pretrained language models; no human participants or animals were involved.

Consent for publication. Not applicable.

Data availability. The WikiText-103 data used in this study are publicly available. Prompt construction procedures and analysis code are provided in the accompanying code repository. Generated continuations and derived evaluation data are available from the corresponding author on reasonable request.

Materials availability. Not applicable.

Code availability. Code and additional reproducibility details are available at https://github.com/maygodwithu/decoding_future_semantic.git.

Author contributions. Jaekeol Choi conceived and designed the study, developed the methodology and software, conducted the experiments, curated and analyzed the data, prepared the visualizations, and wrote and revised the manuscript. The author read and approved the final manuscript.

References

- [1] Belrose, N., Furman, Z., Smith, L., Halawi, D., Ostrovsky, I., McKinney, L., Biderman, S., Steinhardt, J.: Eliciting latent predictions from transformers with the tuned lens. arXiv preprint arXiv:2303.08112 (2023) [arXiv:2303.08112](https://arxiv.org/abs/2303.08112)

- [2] Pal, K., Sun, J., Yuan, A., Wallace, B.C., Bau, D.: Future lens: Anticipating subsequent tokens from a single hidden state. In: Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL), pp. 548–560 (2023). <https://doi.org/10.18653/v1/2023.conll-1.37> . <https://aclanthology.org/2023.conll-1.37/>
- [3] Wu, W., Morris, J.X., Levine, L.: Do language models plan ahead for future tokens? In: First Conference on Language Modeling (2024). <https://openreview.net/forum?id=BaOAvPUyBO>
- [4] Pochinkov, N., Benoit, A., Agarwal, L., Majid, Z.A., Ter-Minassian, L.: Extracting paragraphs from LLM token activations. arXiv preprint arXiv:2409.06328 (2024) [arXiv:2409.06328](https://arxiv.org/abs/2409.06328)
- [5] Dong, Z., Zhou, Z., Liu, Z., Yang, C., Lu, C.: Emergent response planning in LLMs. In: Proceedings of the 42nd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 267, pp. 14301–14320 (2025). <https://proceedings.mlr.press/v267/dong25p.html>
- [6] Pochinkov, N., Volkova, Y., Vasileva, A., Chereddy, S.V.R.: ParaScopes: What do language models activations encode about future text? arXiv preprint arXiv:2511.00180 (2025) [arXiv:2511.00180](https://arxiv.org/abs/2511.00180)
- [7] Hewitt, J., Liang, P.: Designing and interpreting probes with control tasks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 2733–2743 (2019). <https://doi.org/10.18653/v1/D19-1275> . <https://aclanthology.org/D19-1275/>
- [8] Elazar, Y., Ravfogel, S., Jacovi, A., Goldberg, Y.: Amnesic probing: Behavioral explanation with amnesic counterfactuals. Transactions of the Association for Computational Linguistics **9**, 160–175 (2021) <https://doi.org/10.1162/tacl.a.00359>
- [9] Subramani, N., Suresh, N., Peters, M.E.: Extracting latent steering vectors from pretrained language models. In: Findings of the Association for Computational Linguistics: ACL 2022, pp. 566–581 (2022). <https://doi.org/10.18653/v1/2022.findings-acl.48> . <https://aclanthology.org/2022.findings-acl.48/>
- [10] Rimskey, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., Turner, A.M.: Steering llama 2 via contrastive activation addition. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 15504–15522 (2024). <https://doi.org/10.18653/v1/2024.acl-long.828> . <https://aclanthology.org/2024.acl-long.828/>
- [11] Alain, G., Bengio, Y.: Understanding intermediate layers using linear classifier probes. arXiv preprint arXiv:1610.01644 (2016) [arXiv:1610.01644](https://arxiv.org/abs/1610.01644)

- [12] Adi, Y., Kermany, E., Belinkov, Y., Lavi, O., Goldberg, Y.: Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In: International Conference on Learning Representations (2017). <https://openreview.net/forum?id=BJh6Ztuxl>
- [13] Conneau, A., Kruszewski, G., Lample, G., Barrault, L., Baroni, M.: What you can cram into a single $\&\#\ast$ vector: Probing sentence embeddings for linguistic properties. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, pp. 2126–2136 (2018). <https://doi.org/10.18653/v1/P18-1198> . <https://aclanthology.org/P18-1198/>
- [14] Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R.T., Kim, N., Van Durme, B., Bowman, S.R., Das, D., Pavlick, E.: What do you learn from context? probing for sentence structure in contextualized word representations. In: International Conference on Learning Representations (2019). <https://openreview.net/forum?id=SJzSgnRcKX>
- [15] Hewitt, J., Manning, C.D.: A structural probe for finding syntax in word representations. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4129–4138 (2019). <https://doi.org/10.18653/v1/N19-1419> . <https://aclanthology.org/N19-1419/>
- [16] Fayyaz, M., Aghazadeh, E., Modarressi, A., Mohebbi, H., Pilehvar, M.T.: Not all models localize linguistic knowledge in the same place: A layer-wise probing on BERToids’ representations. In: Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, pp. 375–388 (2021). <https://doi.org/10.18653/v1/2021.blackboxnlp-1.29> . <https://aclanthology.org/2021.blackboxnlp-1.29/>
- [17] Pimentel, T., Valvoda, J., Maudslay, R.H., Zmigrod, R., Williams, A., Cotterell, R.: Information-theoretic probing for linguistic structure. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4609–4622 (2020). <https://doi.org/10.18653/v1/2020.acl-main.420> . <https://aclanthology.org/2020.acl-main.420/>
- [18] Voita, E., Titov, I.: Information-theoretic probing with minimum description length. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 183–196 (2020). <https://doi.org/10.18653/v1/2020.emnlp-main.14> . <https://aclanthology.org/2020.emnlp-main.14/>
- [19] Belinkov, Y.: Probing classifiers: Promises, shortcomings, and advances. Computational Linguistics **48**(1), 207–219 (2022) https://doi.org/10.1162/coli_a.00422
- [20] Hernandez, E., Sen Sharma, A., Haklay, T., Meng, K., Wattenberg, M., Andreas, J., Belinkov, Y., Bau, D.: Linearity of relation decoding in transformer language

- models. In: The Twelfth International Conference on Learning Representations (2024). <https://openreview.net/forum?id=w7LU2s14kE>
- [21] Park, K., Choe, Y.J., Veitch, V.: The linear representation hypothesis and the geometry of large language models. In: Proceedings of the 41st International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 235, pp. 39643–39666 (2024). <https://proceedings.mlr.press/v235/park24c.html>
 - [22] nostalgebraist: Interpreting GPT: The Logit Lens. Online. Accessed: Aug. 25, 2026 (2020). <https://www.alignmentforum.org/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens>
 - [23] Yom Din, A., Karidi, T., Choshen, L., Geva, M.: Jump to conclusions: Short-cutting transformers with linear transformations. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pp. 9615–9625 (2024). <https://aclanthology.org/2024.lrec-main.840/>
 - [24] Ghandeharioun, A., Caciularu, A., Pearce, A., Dixon, L., Geva, M.: Patchscopes: A unifying framework for inspecting hidden representations of language models. In: Proceedings of the 41st International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 235, pp. 15466–15490 (2024). <https://proceedings.mlr.press/v235/ghandeharioun24a.html>
 - [25] Jin, C., Rinard, M.: Emergent representations of program semantics in language models trained on programs. In: Proceedings of the 41st International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 235, pp. 22160–22184 (2024). <https://proceedings.mlr.press/v235/jin24e.html>
 - [26] Turner, A.M., Thiergart, L., Leech, G., Udell, D., Vazquez, J.J., Mini, U., MacDiarmid, M.: Steering language models with activation engineering. arXiv preprint arXiv:2308.10248 (2023) [arXiv:2308.10248](https://arxiv.org/abs/2308.10248)
 - [27] Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., Goel, S., Li, N., Byun, M.J., Wang, Z., Mallen, A., Basart, S., Koyejo, S., Song, D., Fredrikson, M., Kolter, J.Z., Hendrycks, D.: Representation engineering: A top-down approach to AI transparency. arXiv preprint arXiv:2310.01405 (2023) [arXiv:2310.01405](https://arxiv.org/abs/2310.01405)
 - [28] Li, K., Patel, O., Viégas, F., Pfister, H., Wattenberg, M.: Inference-time intervention: Eliciting truthful answers from a language model. In: Advances in Neural Information Processing Systems, vol. 36, pp. 41451–41530 (2023). <https://doi.org/10.52202/075280-1797>. https://proceedings.neurips.cc/paper_files/paper/2023/hash/81b8390039b7302c909cb769f8b6cd93-Abstract-Conference.html
 - [29] Todd, E., Li, M.L., Sen Sharma, A., Mueller, A., Wallace, B.C., Bau, D.: Function

- vectors in large language models. In: The Twelfth International Conference on Learning Representations (2024). <https://openreview.net/forum?id=AwyxtyMwaG>
- [30] Arditi, A., Obeso, O., Syed, A., Paleka, D., Panickssery, N., Gurnee, W., Nanda, N.: Refusal in language models is mediated by a single direction. In: Advances in Neural Information Processing Systems, vol. 37 (2024). <https://doi.org/10.52202/079017-4322> . https://proceedings.neurips.cc/paper_files/paper/2024/hash/f545448535dfde4f9786555403ab7c49-Abstract-Conference.html
- [31] Stolfo, A., Balachandran, V., Yousefi, S., Horvitz, E., Nushi, B.: Improving instruction-following in language models through activation steering. In: The Thirteenth International Conference on Learning Representations (2025). <https://openreview.net/forum?id=wozhdnRCtw>
- [32] Zhao, Y., Devoto, A., Hong, G., Du, X., Gema, A.P., Wang, H., He, X., Wong, K.-F., Minervini, P.: Steering knowledge selection behaviours in LLMs via SAE-based representation engineering. In: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 5117–5136 (2025). <https://doi.org/10.18653/v1/2025.naacl-long.264> . <https://aclanthology.org/2025.naacl-long.264/>

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [NPLSupplementary.pdf](#)