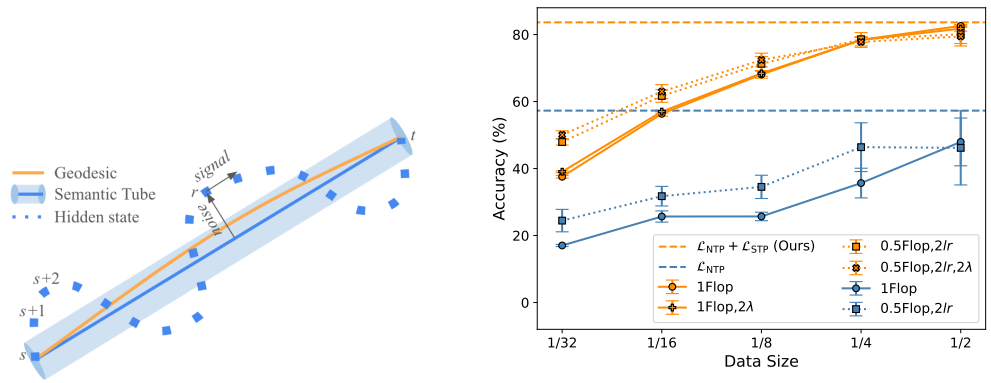
SEMANTIC TUBE PREDICTION: BEATING LLM DATA EFFICIENCY WITH JEPA

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Language Models (LLMs) obey consistent scaling laws—empirical power-law fits that predict how loss decreases with compute, data, and parameters. While predictive, these laws are descriptive rather than prescriptive: they characterize typical training, not optimal training. Surprisingly few works have successfully challenged the data-efficiency bounds implied by these laws—which is our primary focus. To that end, we introduce the Geodesic Hypothesis, positing that token sequences trace geodesics on a smooth semantic manifold and are therefore locally linear. Building on this principle, we propose a novel Semantic Tube Prediction (STP) task, a JEPA-style regularizer that confines hidden-state trajectories to a tubular neighborhood of the geodesic. STP generalizes JEPA to language without requiring explicit multi-view augmentations. We show this constraint improves signal-to-noise ratio, and consequently preserves diversity by preventing trajectory collisions during inference. Empirically, STP allows LLMs to match baseline accuracy with 16× less training data, directly violating the data term of Chinchilla-style scaling laws and demonstrating that principled geometric priors can surpass brute-force scaling.



(a) Semantic Tube (b) Data Efficiency

Figure 1: Semantic Tube improves data efficiency. **(a)** We hypothesize that error-free hidden state trajectories are geodesics, which are locally linear and approximated by the Semantic Tube. The dotted line depicts a trajectory distorted by training loss. Deviations perpendicular to the tube constitute *noise*, while the component along the geodesic represents the *signal*. **(b)** With our approach ($\mathcal{L}_{\text{NTP}} + \mathcal{L}_{\text{STP}}$), accuracy shows a negligible drop when the training dataset is halved, and it matches full-dataset standard fine-tuning ($\mathcal{L}_{\text{NTP}}$) accuracy using only $\frac{1}{16}$ of the training data. In contrast, $\mathcal{L}_{\text{NTP}}$ degrades significantly when the dataset is halved.

1 INTRODUCTION

We argue that empirical scaling laws characterize *typical* rather than *optimal* training, suggesting the rigid power-law barrier is an artifact of current objectives. The core limitation is next-token prediction: a local objective that conflates surface statistical noise with global semantic signal. We

propose a fundamental shift: explicitly constraining hidden state dynamics to separate the error-free semantic trajectory from this noise.

First, we formally demonstrate that, although tokens are discrete, token sequences can be modeled by an Ordinary Differential Equation (ODE). The Picard-Lindelöf Theorem guarantees that, under mild conditions, trajectories originating from distinct initial states will never intersect. In the context of LLMs, if the ODE model holds, this implies that *error-free* generations from distinct prompts maintain their semantic separation, theoretically ruling out mode collapse and preserving diversity.

Next, we hypothesize that the Principle of Least Action is at work. This principle states that the path taken by a system between two points minimizes the "Action" (the integral of the Lagrangian over time), resulting in a "straight line" or geodesic on the underlying manifold. We further hypothesize that, as the manifold is an artifact of the training process, it admits a smooth structure. Consequently, the geodesics are locally linear almost everywhere. In the context of LLMs, this implies that the trajectories of error-free token sequences—and by extension, the trajectories of error-free hidden states—are confined within a tube centered along a straight line.

We designate this structure the **Semantic Tube** (fig. 1) and leverage it to regularize the LLM training process. The Semantic Tube posits that the noise—which causes deviations from the error-free trajectories—concentrates along the directions perpendicular to the tube. Let $s < r < t$ denote the indices of three tokens. We define the *noise* term as $(h_r - h_s)_{\perp h_t - h_s}$, representing the component of $h_r - h_s$ perpendicular to $h_t - h_s$, and the *signal* term as $(h_r - h_s)_{\parallel h_t - h_s}$, representing the component parallel to $h_t - h_s$. Minimizing the noise term is expected to improve the Signal-to-Noise Ratio (SNR) during training. We formulate this as an auxiliary loss term, the Semantic Tube Prediction (STP) loss $\mathcal{L}_{\text{STP}}$, which can be seamlessly integrated into the training objective:

$$\mathcal{L} = \mathcal{L}_{\text{NTP}} + \lambda \cdot \mathcal{L}_{\text{STP}}$$

where $\mathcal{L}_{\text{NTP}}$ is the cross-entropy loss for Next Token Prediction (NTP) and λ is a hyperparameter controlling the strength of the STP loss.

Semantic Tube draws inspiration from the Joint-Embedding Predictive Architecture (JEPA) (Assran et al., 2023; Baevski et al., 2022), which learns to predict the representation of one view based on another. In our approach, we postulate that any segment of a token sequence aligns with the global trajectory; consequently, the predictor reduces to an identity function.

If the Geodesic Hypothesis holds, it entails the following predictions:

- (P1) $\mathcal{L}_{\text{NTP}}$ alone is insufficient for high-quality generation. Consequently, we expect to observe $\mathcal{L}_{\text{NTP}}$ plateau even as $\mathcal{L}_{\text{STP}}$ continues to decrease.
- (P2) Semantic Tube improves SNR, resulting in superior data efficiency (fig. 1) and accuracy.
- (P3) Semantic Tube preserves diversity.
- (P4) We expect to see $\lambda \ll 1$ to accommodate instances where the geodesic deviates from a straight line.
- (P5) The identity function serves as a superior predictor compared to learned projections.

We conducted extensive experiments validating predictions (P1) through (P5). These results provide a strong indication that the Geodesic Hypothesis represents a simplified form of self-consistency for autoregressive sequence models. Furthermore, they confirm the validity of the noise/signal decomposition (fig. 1) and establish Semantic Tube as an effective self-supervised learning objective for LLMs.

2 Training and Inference Dynamics

In this section, we formally analyze training and inference dynamics, proposing that token sequence trajectories can be modeled by an Ordinary Differential Equation (ODE) characterized by ballistic trajectories.

2.1 TRAINING ODE

Let $x_{\leq t}$ denote a token sequence of length t, where x_t represents the t-th token, h_t is the corresponding hidden state, and $f(\cdot)$ denotes the neural network such that $h_t = f(x_{\leq t})$. Each hidden state h_t is subsequently unembedded to predict the next token x_{t+1}.

During training, the predicted token $u(h_t)$ may diverge from the ground truth x_{t+1}; this discrepancy constitutes the training loss. However, due to teacher forcing, we invariably feed the ground truth sequence $x_{\leq t+1}$ into $f(\cdot)$ to generate h_{t+1}. Consequently, assuming a converged network where the loss is minimized, the training dynamics can be modeled as:

$$x_{t+1} = \mathring{u} \circ \mathring{f}(x_{\leq t}) \tag{1}$$

$$h_t = \mathring{f}(x_{\leq t}) + \epsilon_t \tag{2}$$

where $\mathring{f}$ and $\mathring{u}$ represent the functions of the converged network, and ϵ_t denotes the residual unembedding error.

If a time-indexed variable z_t follows the difference equation $z_{t+1} - z_t = g(z_t, t)$, it can be approximated by an ODE of the form $dz_t = g(z_t, t)dt$. While the hidden state dynamics in eq. (2) do not fit this form (as h_{t+1} depends on the entire history $x_{\leq t}$ rather than just h_t), the sequence dynamics in eq. (1) do. Specifically, $x_{\leq t+1} = x_{\leq t} \oplus x_{t+1} = x_{\leq t} \oplus \mathring{u} \circ \mathring{f}(x_{\leq t})$, where $\oplus$ denotes concatenation. Letting $\ominus$ denote the prefix-removal operator, we obtain:

$$x_{\leq t+1} \ominus x_{\leq t} = \mathring{u} \circ \mathring{f}(x_{\leq t}).$$

This formulation closely resembles the update rule $z_{t+1} - z_t = g(z_t, t)$, suggesting that an ODE is a plausible model for the dynamics.

Although tokens are discrete, their embeddings lie in a continuous vector space $x_t \in \mathbb{R}^{d_{\text{model}}}$. Let T denote the maximum sequence length; then the sequence resides in $\mathbb{R}^{T \times d_{\text{model}}}$. In section A, we demonstrate that under specific arrangements, the operation $x_{\leq t+1} \ominus x_{\leq t}$ can be treated as vector subtraction $x_{\leq t+1} - x_{\leq t}$. This leads to the following proposition:

Proposition 1 (Training ODE)**.** *The LLM training process can be modeled as a solution in the token sequence space $\mathbb{R}^{T \times d_{\text{model}}}$ to the ODE:*

$$dx_{\leq t} = \mathring{u} \circ \mathring{f}(x_{\leq t})dt.$$

Proposition 1 models $x_{\leq t}$ as following a ballistic trajectory in $\mathbb{R}^{T \times d_{\text{model}}}$. The Picard-Lindelöf Theorem guarantees that if $\mathring{u} \circ \mathring{f}(\cdot)$ and its partial derivatives with respect to $x_{\leq t}$ are continuous, the ODE admits a unique solution for a given initial condition. Consequently, within this ODE framework, sequences generated from distinct prompts (initial conditions) cannot intersect, theoretically ruling out mode collapse, and preserving diversity.

2.2 MODE COLLAPSE AT INFERENCE TIME

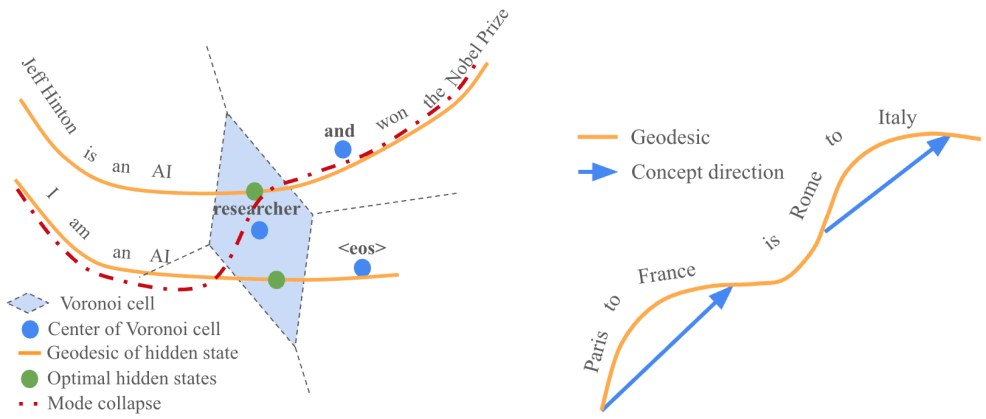
Let h^* denote the optimal trajectory of hidden states, defined as:

$$h_t^* = h_t - \epsilon_t = \mathring{f}(x_{\leq t}) \tag{3}$$

If $\mathring{f}(\cdot)$ is Lipschitz-continuous, then the trajectory h^* is also ballistic.

However, $\mathcal{L}_{\text{NTP}}$ alone may not suffice to drive ϵ_t to zero. Recall that the goal of $\mathcal{L}_{\text{NTP}}$ is to converge $u(h_t)$ to x_{t+1}. Since the hidden state h_t is continuous while the token x_{t+1} is discrete, the training process can be modeled as finding the correct Voronoi cell, without stipulating the exact location within the cell. This flexibility is necessary for the Picard-Lindelöf Theorem to apply: as illustrated in fig. 2, it allows error-free geodesics (h_t^*) to traverse the same Voronoi cell at distinct locations, thereby avoiding intersection. Nevertheless, h_t may drift onto an incorrect geodesic within the cell, leading to mode collapse.

This analysis indicates that $\mathcal{L}_{\text{NTP}}$ alone is insufficient for generation quality, strongly motivating an additional loss term ($\mathcal{L}_{\text{STP}}$) to explicitly minimize ϵ_t. It also implies that within the correct Voronoi cell, $\mathcal{L}_{\text{NTP}}$ may plateau while $\mathcal{L}_{\text{STP}}$ continuously decreases. Therefore, (P1).



(a) Voronoi cell (b) Geodesic alignment

Figure 2: **(a)** Two hidden state trajectories with similar prefixes pass through the Voronoi cell of the "**researcher**" token at different locations, leading to different next hidden states and hence different next tokens. Since $\mathcal{L}_{\text{NTP}}$ cannot guarantee that h_t converges to h_t^* (optimal hidden state), h_t can be misplaced on another geodesic. This leads to mode collapse (the red dotted line mistakenly continues the generation, misattributing Hinton's Nobel Prize to an arbitrary person, or if the error deviates in the opposite direction and precludes a winner). **(b)** When the sentence aligns on a geodesic, the concept direction naturally aligns.

In section B, we demonstrate that in the infinite-width limit (Yang & Littwin, 2021), the inference process can be modeled as a Stochastic Differential Equation (SDE) with a Brownian motion term.

3 Semantic Tube Prediction

A key challenge in minimizing the error ϵ_t is that the optimal trajectory h^* remains latent and unknown. To address this, we must postulate a structural property that allows us to estimate h^*, leading us to the Geodesic Hypothesis. In this section, we formalize this hypothesis and subsequently introduce Semantic Tube Prediction (STP).

3.1 Semantic Tube

If the Principle of Least Action holds, the trajectories of the token sequence $x_{\leq t+1}$ in eq. (1) must be geodesics, which are locally linear almost everywhere. Since $h_t^* = \mathring{f}(x_{\leq t})$, when $\mathring{f}(\cdot)$ is smooth enough, h_t^* is also expected to be locally linear almost everywhere. Hence the Geodesic Hypothesis:

The trajectory of $x_{\leq t} \in \mathbb{R}^{T \times d_{\text{model}}}$ is locally linear almost everywhere. Similarly, the trajectory $h_t - \epsilon_t \in \mathbb{R}^d$ is locally linear almost everywhere.

We first formally define local linearity. Subsequently, we demonstrate that the Semantic Tube compresses the trajectory h_t within a tube centered at h_t^*.

Definition 1 (Local Linearity)**.** *A time-indexed trajectory h^* is defined as locally linear if $\exists \tau, \exists \varepsilon$ such that for any time indices $s < r < t$ satisfying $|t - s| \leq \tau$, we have:*

$$\|(h_r^* - h_s^*)_{\perp h_t^* - h_s^*}\|_2 \leq \varepsilon \tag{4}$$

where $x_{\perp y}$ denotes the component of vector x that is perpendicular to vector y.

Definition 1 captures the intuition that if a trajectory is locally linear, each local segment can be approximated by a straight line connecting its endpoints.

Next, we demonstrate that the Semantic Tube forces h to approximate h^*.

Lemma 1 (Straightening Lemma)**.** *If $h_s = h_s^*$, $h_t = h_t^*$, and $\mathcal{L}_{\text{STP}} \leq \epsilon$ for all r satisfying $s < r < t$, then*

$$\|(h_r - h_s)_{\perp h_t^* - h_s^*}\|_2 \leq \sqrt{2\epsilon}\|h_r - h_s\|_2.$$

Proof is deferred to section D.

Let $\|h_r - h^*\|_2 = \min_{r'} \|h_r - h^*_{r'}\|_2$ denote the minimum distance from h_r to the trajectory h^*. We establish the following theorem:

Theorem 1 (Semantic Tube)**.** *If h^* is locally linear and for all r satisfying $0 \leq s < r < t \leq \tau$, $\mathcal{L}_{\text{STP}} \to 0$, then*

$$\|h_r - h^*\|_2 \lesssim \varepsilon$$

Proof is deferred to section E.

In practice, the indices $s < r < t$ are selected randomly. Consequently, minimizing $\mathcal{L}_{\text{STP}}$ effectively drives $\mathbb{E}[1 - \cos(h_t - h_r, h_r - h_s)] \to 0$. By Markov's inequality, for any ϵ, $P(1 - \cos(h_t - h_r, h_r - h_s) > \epsilon) \to 0$. This leads to the following corollary:

Corollary 1 (Random Tube)**.** *For randomly selected $s < r < t$, if $\mathcal{L}_{\text{STP}} \to 0$, then for any ϵ,*

$$P(\|h_r - h^*\|_2 > \varepsilon + \epsilon) \to 0$$

Corollary 1 implies that if $\mathcal{L}_{\text{STP}} \to 0$ for a given sequence, then with high probability, the trajectory of the sequence's hidden states is confined within a tube centered around the optimal trajectory h^*.

However, at inference time, the Brownian motion term diverges into a cone whose radius scales as $\propto \sigma_t \sqrt{t}$, see section F for details.

3.2 Practical Considerations

Since the forward pass naturally computes h_s, h_r, and h_t, the STP loss introduces negligible computational overhead—primarily the cost of computing cosine similarity. This is significantly more efficient than the fractional extra forward passes required by LLM-JEPA (Huang et al., 2025). Furthermore, because indices s, r, and t can be selected randomly, STP eliminates the need for manual scaffolding of a two-view structure. In summary, STP effectively addresses the two primary limitations that have hindered the broader adoption of LLM-JEPA. Additionally, STP avoids the complexity of a predictor network (often a requirement in LLM-JEPA), as local linearity implies an identity predictor. Like LLM-JEPA, the STP loss is applied exclusively during training and is not required at inference time.

Further implementation details are provided in section G.

3.3 Related Work

JEPAs (Assran et al., 2023; Baevski et al., 2022) learn predictive representations across views, offering theoretical benefits Littwin et al. (2024) despite the risk of dimensional collapse Jing et al. (2021); Kenneweg et al. (2025). While recent works extend these objectives to LLMs Barrault et al. (2024); Wang & Sun (2025), LLM-JEPA (Huang et al., 2025) is bottlenecked by manual two-view scaffolding and the computational cost of additional forward passes, neither is a problem for $\mathcal{L}_{\text{STP}}$.

Scaling Laws govern the power-law relationship between compute, data, and parameters in both pre-training (Kaplan et al., 2020; Hoffmann et al., 2022) and fine-tuning (Zhang et al., 2024). While recent data efficiency research emphasizes identifying high-density subsets (Sorscher et al., 2022) or synthetic curation (Gunasekar et al., 2023; Muennighoff et al., 2023), $\mathcal{L}_{\text{STP}}$ enhances the training SNR directly, obviating the need for explicit data subset selection.

SDE/ODE Perspective: Kong et al. (2020) interpreted ResNets as "Neural SDEs" which has a Brownian motion term. While Tong et al. (2025) recently adapted ODEs for LLMs, they model evolution across network depth (layers). Our work takes an orthogonal approach, focusing instead on the temporal dynamics of hidden states across the token sequence.

The Linear Representation Hypothesis (LRH) (Park et al., 2024; 2025) posits that simple concepts are encoded as directions in the representation space, whereas the Geodesic Hypothesis suggests that both simple and composed concepts (expressed as token sequences) follow locally linear trajectories.

Consequently, the vector arithmetic observed in LRH ($\vec{v}_{Paris} - \vec{v}_{France} + \vec{v}_{Italy} \approx \vec{v}_{Rome}$) emerges naturally from path linearity ($\vec{v}_{Paris}, \vec{v}_{to}, \vec{v}_{France}, \vec{v}_{is}, \vec{v}_{Rome}, \vec{v}_{to}, \vec{v}_{Italy}$ aligns on almost a straight line, see fig. 2).

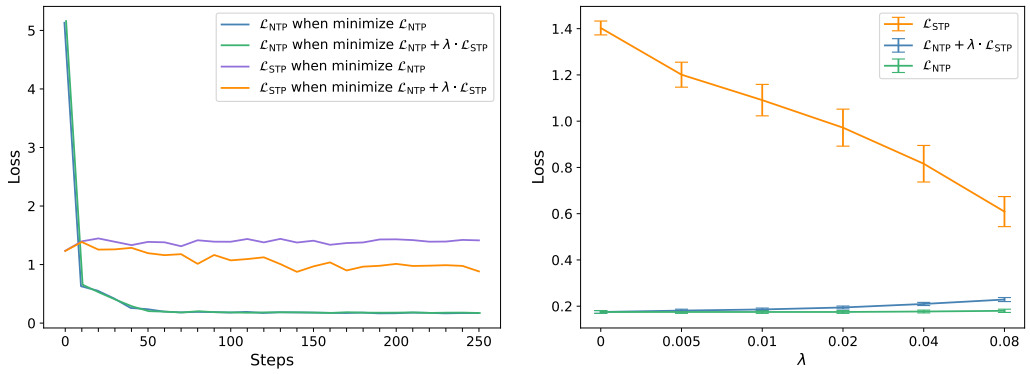
The application of **geodesic geometry to LLMs** remains underexplored, with existing studies primarily restricted to interpolating representations across models (Deng et al., 2025; Yu et al., 2024).

See section I for more related work.

4 Experiments

Implementing $\mathcal{L}_{\text{STP}}$ is straightforward with HuggingFace `transformers`. When computing loss, we grab per-token `hidden_state` h from last layer, pick (random) indices $s < r < t$, and compute $1 - \cos(h_t - h_r, h_r - h_s)$. Across all experiments, we follow LLM-JEPA (Huang et al., 2025) to pick 5 random seeds: 82, 23, 37, 84, and 4, and report both mean accuracy and standard deviation. This also allows us to report p-value of paired, single-tailed t-Test. We inherit optimal number of epochs and learning rate from LLM-JEPA. λ is separately tuned.

4.1 Loss Landscape



(a) Loss curve

(b) Loss vs. λ

Figure 3: Loss landscape. **(a)** When $\mathcal{L}_{\text{NTP}}$ plateaus, $\mathcal{L}_{\text{STP}}$ continues to decrease. Furthermore, minimizing $\mathcal{L}_{\text{NTP}}$ does not automatically minimize $\mathcal{L}_{\text{STP}}$. **(b)** Across a wide range of λ, increasing λ on a logarithmic scale reduces $\mathcal{L}_{\text{STP}}$ linearly, while $\mathcal{L}_{\text{NTP}}$ remains unchanged.

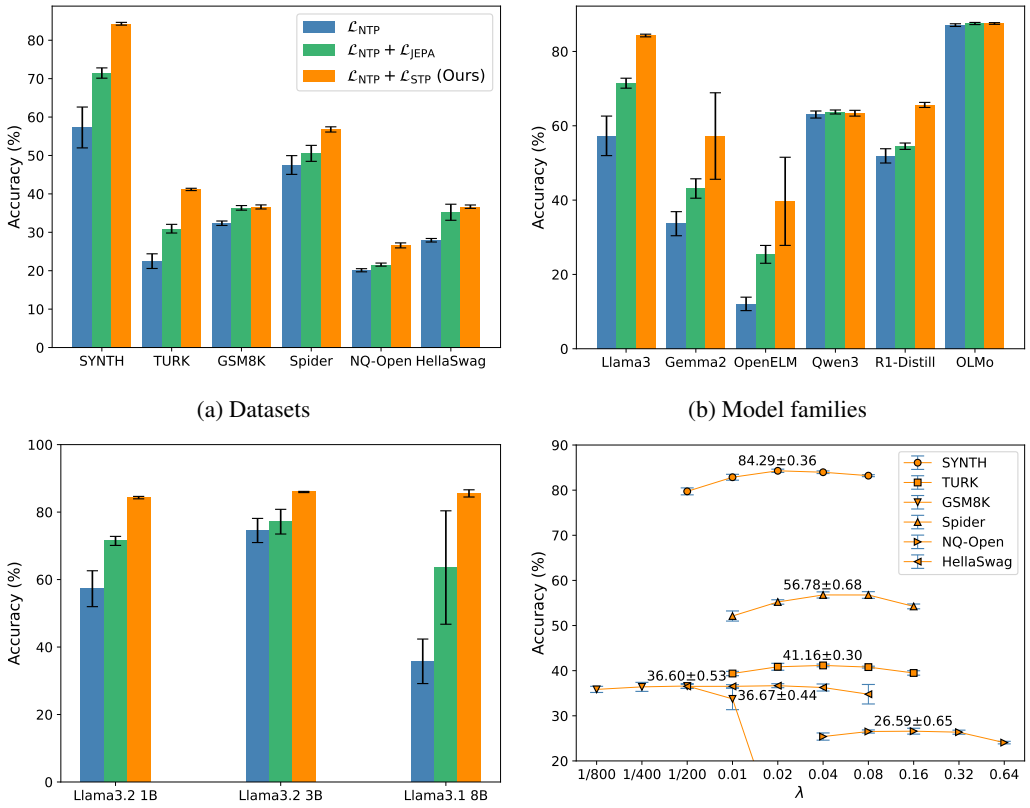
We begin by analyzing the loss landscape by fine-tuning Llama-3.2-1B-Instruct (Grattafiori et al., 2024) on the NL-RX-SYNTH (Locascio et al., 2016) dataset.

Figure 3(a) demonstrates that in regular fine-tuning, minimizing $\mathcal{L}_{\text{NTP}}$ does not automatically minimize $\mathcal{L}_{\text{STP}}$. With the Semantic Tube, however, $\mathcal{L}_{\text{STP}}$ continues to decrease even after $\mathcal{L}_{\text{NTP}}$ plateaus, corroborating (P1). Moreover, while $\mathcal{L}_{\text{NTP}}$ remains comparable between regular and Semantic Tube fine-tuning, there is a significant gap in $\mathcal{L}_{\text{STP}}$. This confirms that the SNR gain is driven by $\mathcal{L}_{\text{STP}}$, validating the analysis in section 2.2 that $\mathcal{L}_{\text{NTP}}$ alone is insufficient for generation quality and that $\mathcal{L}_{\text{STP}}$ acts as a necessary complement.

Figure 3(b) illustrates that increasing λ on a logarithmic scale reduces $\mathcal{L}_{\text{STP}}$ linearly across a wide range, while $\mathcal{L}_{\text{NTP}}$ remains stable. Given $\mathcal{L}_{\text{STP}} = 1 - \cos(h_t - h_r, h_r - h_s)$, a value of $\mathcal{L}_{\text{STP}} > 1.0$ implies that the trajectory vector $h_t - h_r$ diverges significantly (essentially reversing direction) relative to $h_r - h_s$. At $\lambda = 0$ (regular fine-tuning), $\mathcal{L}_{\text{STP}} \approx 1.4$ indicates a trajectory resembling erratic Brownian motion. At $\lambda = 0.08$, $\mathcal{L}_{\text{STP}}$ drops to 0.6, reflecting a substantially smoother path. Notably, while the optimal performance is achieved at $\lambda = 0.02$ (table 3), the accuracy at $\lambda = 0.08$ is only marginally lower (fig. 4).

4.2 BETTER ACCURACY

On Various Datasets: We first fine-tune Llama-3.2-1B-Instruct to demonstrate that Semantic Tube yields significant accuracy improvements over regular fine-tuning and LLM-JEPA across diverse datasets: NL-RX-SYNTH, NL-RX-TURK (Locascio et al., 2016), GSM8K (Cobbe et al., 2021), Spider (Yu et al., 2018), NQ-Open (Lee et al., 2019), and HellaSwag (Zellers et al., 2019). Figure 4(a) illustrates the superior performance of Semantic Tube compared to regular fine-tuning and LLM-JEPA.



(a) Datasets

(b) Model families

(c) Model sizes

(d) λ vs. accuracy

Figure 4: Semantic Tube ($\mathcal{L}_{\text{NTP}} + \mathcal{L}_{\text{STP}}$, our approach) demonstrates superior performance across **(a)** datasets, **(b)** model families, and **(c)** model sizes compared to regular fine-tuning ($\mathcal{L}_{\text{NTP}}$) and LLM-JEPA ($\mathcal{L}_{\text{NTP}}+\mathcal{L}_{\text{JEPA}}$). **(d)** Impact of λ tuning on Llama-3 1B across various datasets. In most cases, peak performance is achieved within the range of 0.01 to 0.08.

On Various Model Families: Next, we extend our evaluation to various model families. In addition to Llama, we evaluate gemma-2-2b-it (Team et al., 2024), OpenELM-1_1B-Instruct (Mehta et al., 2024), and OLMo-2-0425-1B-Instruct (OLMo et al., 2024) on NL-RX-SYNTH, as well as Qwen3-1.7B (Yang et al., 2025) and DeepSeek-R1-Distill-Qwen-1.5B (DeepSeek-AI et al., 2025) on GSM8K. The results are presented in fig. 4(b).

On Various Model Sizes: Finally, we examine scalability across model sizes using Llama-3 1B, 3B, and 8B models. Results are shown in fig. 4(c).

4.3 DATA EFFICIENCY

Data efficiency is another crucial metric demonstrating improved SNR. We randomly select subsets of $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$, and $\frac{1}{32}$ of the NL-RX-SYNTH dataset and perform both Semantic Tube and regular fine-tuning on Llama-3 1B, 3B, and 8B models. To compensate for the reduced number of training steps, we scale the epochs proportionally: with a $\frac{1}{n}$ dataset fraction, we run $n\times$ epochs. For Semantic

Tube, accuracy shows a negligible drop when the training dataset is halved and remains robust until the dataset is reduced to $\frac{1}{16}$, at which point it matches the accuracy of regular fine-tuning on the full dataset. In contrast, regular fine-tuning suffers a significant drop immediately when the dataset is halved. See fig. 1 for 1B results and fig. 8 for 3B and 8B results.

We also experimented with half compute ($\frac{n}{2}\times$ epochs) combined with a $2\times$ learning rate. In both full and half compute scenarios, we also tested $2 \times \lambda$. Interestingly, although the half-compute, double-learning-rate setting does not yield optimal accuracy at $\frac{1}{2}$ or full training data, it performs comparatively better when the dataset fraction is $< \frac{1}{2}$.

The improved accuracy and data efficiency provide strong evidence that Semantic Tube improves SNR (see section H for formal proofs linking SNR to accuracy and data efficiency). This validates (P2) and supports the proposed noise/signal decomposition in fig. 1, where the component perpendicular to the tube represents noise. Consequently, it supports the hypothesis that the geodesic is locally linear; otherwise, it could not be effectively approximated by the tube.

4.4 Preserving Diversity

In this section, we demonstrate that Semantic Tube preserves diversity. In the NL-RX-SYNTH dataset, some regular expressions end with "$.*$", while others end with "$.*.*$". Although functionally equivalent, these variations represent a nuanced preference by the dataset creator; a robust neural network should be able to learn and preserve this diversity. As shown in table 1, we find that regular fine-tuning struggles to learn either pattern effectively. LLM-JEPA learns the former pattern well but fails on the latter, likely because the former dominates the training set by a factor of $35\times$. In contrast, Semantic Tube successfully learns both patterns. We list representative samples from the SYNTH dataset ending with either "$.*$" or "$.*.*$" in table 2.

Table 1: Accuracy on functionally equivalent regular expression suffixes "$.*$" and "$.*.*$". Semantic Tube effectively captures nuanced preferences, whereas LLM-JEPA exhibits mode collapse by biasing towards "$.*$", which is $35\times$ more prevalent in the training set than "$.*.*$".

Suffix	Semantic Tube	Regular	LLM-JEPA
$.*$	88.5%	29.9%	68.9%
$.*.*$	68.0%	28.0%	32.0%

Following LLM-JEPA, we compute the singular value decomposition (SVD) of $\mathrm{Enc}(\mathrm{Text}) - \mathrm{Enc}(\mathrm{Code})$, which exhibits polymorphism. We conjecture that this mechanism allows Semantic Tube to maintain flexibility and preserve diversity. See section L for details.

Collectively, these results validate (P3).

See sections M and N for how to tune λ and ablation study. We note that most optimal λ is between 0.01 and 0.08 (table 3), which validates (P4). Also, the **Pred** variant—which trains a linear projector P to minimize $\mathcal{L}_{\mathrm{STP}} = 1 - \cos(P(h_r - h_s), h_t - h_r)$—results in degraded performance in all configurations. This validates (P5).

5 Conclusion

This paper proposes the Geodesic Hypothesis, which posits that token sequence trajectories on the LLM manifold are locally linear geodesics. Based on it, we introduce Semantic Tube Prediction (STP)—a learning objective complementary to Next Token Prediction—which compresses hidden state trajectories into a signal-rich tube centered on the geodesic. Our approach generalizes LLM-JEPA by eliminating the need for manual scaffolding of two-view structures, additional compute, or auxiliary predictors. Empirically, STP significantly improves Signal-to-Noise Ratio, allowing models to maintain accuracy even when training data is reduced to $\frac{1}{16}$, thereby challenging standard Power Law scaling. Our framework unifies the Linear Representation and Manifold Hypotheses under the Principle of Least Action.

REFERENCES

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International conference on machine learning*, pp. 1298–1312. PMLR, 2022.

Loïc Barrault, Paul-Ambroise Duquenne, Maha Elbayad, Artyom Kozhevnikov, Belen Alastruey, Pierre Andrews, Mariano Coria, Guillaume Couairon, Marta R Costa-jussà, David Dale, et al. Large concept models: Language modeling in a sentence representation space. *arXiv preprint arXiv:2412.08821*, 2024.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. *Advances in neural information processing systems*, 28, 2015.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, John Schulman, Jacob Hilton, Melanie Knight, Adrian Weller, Dario Amodei, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications. Wiley-Interscience, 1991. ISBN 0-471-06259-6.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Chenhui Deng, Yunsheng Bai, and Haoxing Ren. Chipalign: Instruction alignment in large language models for chip design via geodesic interpolation. In *2025 62nd ACM/IEEE Design Automation Conference (DAC)*, pp. 1–7. IEEE, 2025.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 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. 4171–4186, 2019.

Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*, 2023.

Olivier J. Hénaff, Yoon Bai, Julie A. Charlton, Ian Nauhaus, Eero P. Simoncelli, and Robbe L. T. Goris. Primary visual cortex straightens natural video trajectories. *Nature Communications*, 12(1):5982, oct 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-25939-z. URL https://doi.org/10.1038/s41467-021-25939-z.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Eghbal Hosseini and Evelina Fedorenko. Large language models implicitly learn to straighten neural sentence trajectories to construct a predictive representation of natural language. *Advances in Neural Information Processing Systems*, 36:43918–43930, 2023.

Jiri Hron, Yasaman Bahri, Jascha Sohl-Dickstein, and Roman Novak. Infinite attention: Nngp and ntk for deep attention networks. In *International Conference on Machine Learning*, pp. 4376–4386. PMLR, 2020.

Hai Huang, Yann LeCun, and Randall Balestriero. Llm-jepa: Large language models meet joint embedding predictive architectures. *arXiv preprint arXiv:2509.14252*, 2025.

Ferenc Huszár. How (not) to train your generative model: Scheduled sampling, likelihood, adversary? *arXiv preprint arXiv:1511.05101*, 2015.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. *arXiv preprint arXiv:2110.09348*, 2021.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Tristan Kenneweg, Philip Kenneweg, and Barbara Hammer. Jepa for rl: Investigating joint-embedding predictive architectures for reinforcement learning. *arXiv preprint arXiv:2504.16591*, 2025.

Bobak Kiani, Jason Wang, and Melanie Weber. Hardness of learning neural networks under the manifold hypothesis. *Advances in Neural Information Processing Systems*, 37:5661–5696, 2024.

Taeuk Kim, Kang Min Yoo, and Sang-goo Lee. Self-guided contrastive learning for BERT sentence representations. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 2528–2540, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.197. URL https://aclanthology.org/2021.acl-long.197/.

Lingkai Kong, Jimeng Sun, and Chao Zhang. Sde-net: Equipping deep neural networks with uncertainty estimates. In *37th International Conference on Machine Learning, ICML 2020*, pp. 5361–5371. International Machine Learning Society (IMLS), 2020.

Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, Fujie Huang, et al. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. Latent retrieval for weakly supervised open domain question answering. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6086–6096, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1612. URL https://aclanthology.org/P19-1612/.

Etai Littwin, Omid Saremi, Madhu Advani, Vimal Thilak, Preetum Nakkiran, Chen Huang, and Joshua Susskind. How jepa avoids noisy features: The implicit bias of deep linear self distillation networks. *Advances in Neural Information Processing Systems*, 37:91300–91336, 2024.

Nicholas Locascio, Karthik Narasimhan, Eduardo DeLeon, Nate Kushman, and Regina Barzilay. Neural generation of regular expressions from natural language with minimal domain knowledge. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1918–1923, 2016.

Sachin Mehta, Mohammad Hossein Sekhavat, Qingqing Cao, Maxwell Horton, Yanzi Jin, Chenfan Sun, Iman Mirzadeh, Mahyar Najibi, Dmitry Belenko, Peter Zatloukal, et al. Openelm: An efficient language model family with open training and inference framework. *arXiv preprint arXiv:2404.14619*, 2024.

Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A Raffel. Scaling data-constrained language models. *Advances in Neural Information Processing Systems*, 36:50358–50376, 2023.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, et al. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*, 2024.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=bVTM2QKYuA.

Michael Robinson, Sourya Dey, and Tony Chiang. Token embeddings violate the manifold hypothesis. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Mariia Seleznova and Gitta Kutyniok. Neural tangent kernel beyond the infinite-width limit: Effects of depth and initialization. In *International Conference on Machine Learning*, pp. 19522–19560. PMLR, 2022.

C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3): 379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x.

HT Siegleman and ED Sontag. On the computational power of neural networks. *Journal of Computer and System Sciences*, 50:132–150, 1995.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. *Advances in Neural Information Processing Systems*, 35:19523–19536, 2022.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Anh Tong, Thanh Nguyen-Tang, Dongeun Lee, Duc Nguyen, Toan Tran, David Hall, CHEONG-WOONG KANG, and Jaesik Choi. Neural ode transformers: Analyzing internal dynamics and adaptive fine-tuning. In *International Conference on Learning Representations (ICLR)*. International Conference on Learning Representations, 2025.

Boshi Wang and Huan Sun. Is the reversal curse a binding problem? uncovering limitations of transformers from a basic generalization failure. *arXiv preprint arXiv:2504.01928*, 2025.

Nick Whiteley, Annie Gray, and Patrick Rubin-Delanchy. Statistical exploration of the manifold hypothesis. *Journal of the Royal Statistical Society: Series B*, 2025.

Ronald J Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280, 1989.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Greg Yang and Edward J Hu. Tensor programs iv: Feature learning in infinite-width neural networks. In *International Conference on Machine Learning*, pp. 11727–11737. PMLR, 2021.

Greg Yang and Etai Littwin. Tensor programs iib: Architectural universality of neural tangent kernel training dynamics. In *International conference on machine learning*, pp. 11762–11772. PMLR, 2021.

Hanlin Yu, Berfin Inal, and Marco Fumero. Connecting neural models latent geometries with relative geodesic representations. In *NeurIPS 2024 Workshop on Symmetry and Geometry in Neural Representations*, 2024.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, et al. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

Biao Zhang, Zhongtao Liu, Colin Cherry, and Orhan Firat. When scaling meets llm finetuning: The effect of data, model and finetuning method. In *The Twelfth International Conference on Learning Representations*, 2024.

A TRAINING ODE

In this section, we present a form of $u(\cdot)$ and $f(\cdot)$ such that $x_{\leq t+1} \ominus x_{\leq t} = x_{\leq t+1} - x_{\leq t}$. Throughout the section, we slightly abuse notation by letting x_t denote both a token and its embedding vector $x_t \in \mathbb{R}^{d_{\text{model}}}$, and letting $x_{\leq t}$ denote both a token sequence and its embedding vector $x_{\leq t} \in \mathbb{R}^{T \times d_{\text{model}}}$:

$$x_{\leq t} = [x_1, ..., x_t, 0, ..., 0].$$

Let $f(x_{\leq t}) \in \mathbb{R}^d$. Let $u(\cdot) : \mathbb{R}^d \to \mathbb{R}^{d_{\text{model}}}$ be the unembedding function that maps the hidden state back to the token embedding.

Note that we need a function to lift $u(f(x_{\leq t}))$ from $\mathbb{R}^{d_{\text{model}}}$ to $\mathbb{R}^{T \times d_{\text{model}}}$. Define $v(\cdot, \cdot) : \mathbb{R}^{d_{\text{model}}} \times \mathbb{N} \to \mathbb{R}^{T \times d_{\text{model}}}$ such that

$$v(x, t) = [0, ..., 0, \underbrace{x}_{\text{index } t+1}, 0, ..., 0]$$

Hence, we have

$$x_{\leq t+1} = v(u(f(x_{\leq t})), t)$$

Define the $\ominus$ operator as

$$x_{\leq t+1} \ominus x_{\leq t} = v(x_{t+1}, t)$$

By the definition of $v(\cdot, \cdot)$, we have

$$x_{\leq t+1} \ominus x_{\leq t} = x_{\leq t+1} - x_{\leq t}$$

Note that the network is now in the form $v(u(f(x_{\leq t})), t)$, which can be written as $g(x_{\leq t}, t)$ and satisfies the formulation of an ODE.

B INFERENCE SDE

At training time, the unembedding error ϵ_t does not propagate to the next token. However, at inference time, h_{t+1} depends (indirectly) on h_t, causing ϵ_t to accumulate into a Brownian motion term.

Yang & Littwin (2021) established that in the limit of infinite width, the pre-activations of a neural network (and thus the hidden state) are well-approximated by Gaussian processes. Hence, we can assume ϵ_t are i.i.d. Gaussian. Furthermore, as shown by (Yang & Littwin, 2021), ϵ_t remains i.i.d. Gaussian when passed through a randomly initialized neural network, which remains constant in the infinite-width limit. Consequently, ϵ_t accumulates to form a Brownian motion term dW_t. Thus the inference process can be modeled by a Stochastic Differential Equation (SDE).

Proposition 2 (Inference SDE)**.** *The inference process of an LLM can be modeled by an SDE in the token sequence space* $\mathbb{R}^{T \times d_{\text{model}}}$*,*

$$dx_{\leq t} = \mathring{u} \circ \mathring{f}(x_{\leq t})dt + \sigma_t dW_t$$

Consider the example in fig. 2, if the Brownian motion shifts the top trajectory to the bottom, mode collapse occurs. Conversely, if the bottom trajectory shifts to the top, mode collapse occurs. This motivates the construction of an approach to explicitly suppress ϵ_t. Indeed, section 4.1 demonstrates that next token prediction alone is insufficient for high-quality generation, making our approach a necessary complement.

C CONTEXT-AWARE HIDDEN STATE

We can view $h_t - h_s$ as the semantic evolution induced by the sub-sequence $x_{[s,t]}$ given the context $x_{\leq s}$. In this sense, $h_t - h_s$ acts as a context-aware hidden state transition, which is significantly more informative than the static hidden state of the isolated sub-sequence $x_{[s,t]}$.

For example, given the prefix $\vec{v}_{\text{The}}, \vec{v}_{\text{capital}}, \vec{v}_{\text{of}}$, appending the token $\vec{v}_{\text{France}}$ shifts the overall semantic trajectory toward $\vec{v}_{Paris}$. However, given a different prefix $\vec{v}_{\text{The}}, \vec{v}_{\text{language}}, \vec{v}_{\text{of}}$, appending the same token $\vec{v}_{France}$ shifts the trajectory toward $\vec{v}_{French}$. If we were to compute the hidden state of $\vec{v}_{France}$ in isolation, we would lose this contextual nuance and fail to capture the context-specific semantic semantic shift.

Thus, $h_t - h_s$ serves as a context-aware representation of the added information.

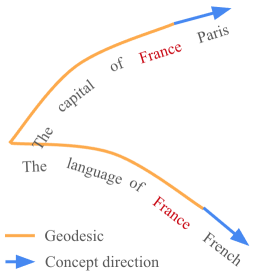


Figure 5: The same token $\vec{v}_{France}$ directs the geodesic along different concept directions when appended to distinct prefixes, illustrating the necessity of the context-aware state difference $h_t - h_s$.

D Proof of the Straightening Lemma

In this section, we provide the proof for lemma 1. The objective is to show

$$\|(h_r - h_s)_{\perp h_t^* - h_s^*}\|_2 \leq \sqrt{2\epsilon}\|h_r - h_s\|_2.$$

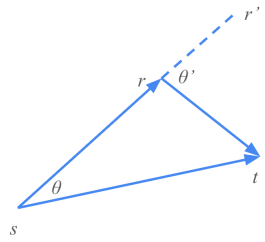


Figure 6: Geometric illustration for the proof of lemma 1

Referring to fig. 6, we have

$$\|(h_r - h_s)_{\perp h_t^* - h_s^*}\|_2 = \|h_r - h_s\|_2 \cdot \sin\theta$$

Since $\theta' \geq \theta$, it follows that

$$\|(h_r - h_s)_{\perp h_t^* - h_s^*}\|_2 \leq \|h_r - h_s\|_2 \cdot \sin\theta'$$

We also have

$$\mathcal{L}_{\text{STP}} = 1 - \cos\theta' \leq \epsilon$$

When ϵ is sufficiently small, we can approximate $\cos\theta' \approx 1 - \frac{\theta'^2}{2}$. Hence

$$\frac{\theta'^2}{2} \lesssim \epsilon$$

Rearranging gives

$$\theta' \lesssim \sqrt{2\epsilon}$$

Also, when θ' is sufficiently small, $\sin\theta' \approx \theta'$. Therefore

$$\|(h_r - h_s)_{\perp h_t^* - h_s^*}\|_2 \leq \|h_r - h_s\|_2 \cdot \sin\theta' \lesssim \sqrt{2\epsilon}\|h_r - h_s\|_2. \qquad \square$$

E Proof of the Semantic Tube Theorem

Proof of theorem 1. First prove for the case $h_s = h_s^*$ and $h_t = h_t^*$. In this scenario, $\|h_r - h_s\|_2 = \|h_r - h_s^*\|$. Applying the triangle inequality yields $\|h_r - h_s^*\| \leq \|h_r^* - h_s^*\|_2 + \epsilon_r$. Notice h_r^* and h_s^* are fixed, by lemma 1, $\|(h_r - h_s)_{\perp h_t^* - h_s^*}\|_2 \to 0$. By definition 1 and the triangle inequality, it follows that $\|h_r - h^*\|_2 \lesssim \varepsilon$

In LLMs, it is standard to assume all sequences begin with <bos> and end with <eos>; thus, it is reasonable to assume the boundary conditions $h_0 = h_0^*$ and $h_\tau = h_\tau^*$.

We introduce two auxiliary tokens, <before-bos> and <after-eos>. The token <before-bos> appears only at the 0-th position and always precedes <bos>, while <after-eod> appears only at the $\tau + 1$-th position and always follow s<eos>. This augmentation increases the total sequence length from τ to $\tau+2$. By anchoring the sequence with <before-bos> and <after-eos>, we ensure that the boundary conditions $h_0 = h_0^*$ and $h_{\tau+1} = h_{\tau+1}^*$ are satisfied.

The proof follows from these conditions. □

F Inference Cone

As STP explicitly reduces ϵ_t, it lowers σ_t in the Brownian Motion term of proposition 2. At inference time, the Brownian motion term causes the token sequence trajectory diverge into a cone whose radius grows at a rate $\propto \sigma_t\sqrt{t}$. A lower σ_t reduces the probability that the cone collides with another token sequence, which would causes mode collapse (fig. 7).

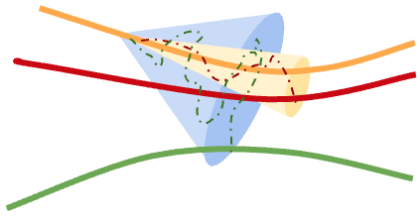


Figure 7: The inference cone defines the probabilistic range of a Brownian motion, and its radius grows $\propto \sigma_t\sqrt{t}$. A larger σ_t leads to a wider cone, which has a high probability of colliding with a token sequence trace that is far away (blue cone and green geodesic), while a smaller σ_t leads to a narrower cone that may only collide with a nearby trace (yellow cone and red geodesic). The dotted red and green fine lines are the Brownian motions confined by the yellow and blue cones, respectively.

Proposition 3 (Inference Cone)**.** *The distortion between h_t and h_t^* behaves as a Gaussian process, where the scale of the deviation grows as $\|h_t - h_t^*\|_2 \propto \sigma\sqrt{t}$*

Proof. According to proposition 2, at inference time, we model the token sequence trajectory as following an SDE $dx_{\leq t} = \mathring{u} \circ \mathring{f}(x_{\leq t})dt + \sigma_t dW_t$, where $\sigma_t dW_t$ is a Brownian motion. Let $h_t = \mathring{f}(x_{\leq t})$ be the hidden state. Let $x^*_{\leq t}$ be the error-free generation satisfying $dx^*_{\leq t} = \mathring{u} \circ \mathring{f}(x^*_{\leq t})dt$, and let $h_t^* = \mathring{f}(x^*_{\leq t})$ be the error-free hidden state. We can quantify the distortion between h_t and h_t^* by examining how the Brownian motion is transformed by $\mathring{f}$.

Yang & Littwin (2021) establishes that in the infinite-width limit, $\mathring{f}$ converges to a Neural Tangent Kernel (NTK) determined by random initialization. It further showed that Gaussian noise remains Gaussian when passed through a randomly initialized network. Hence, a Brownian motion remains a Brownian motion when passed through $\mathring{f}$. Therefore,

$$h_t - h_t^* = \sum_{s \leq t} \epsilon_s$$

where ϵ_s are Gaussian noises. By Donsker's theorem, when $t \to \infty$, $\frac{1}{\sqrt{t}} \sum_{s \leq t} \epsilon_s \sim N(0, \Sigma)$. Consequently, the magnitude of the distortion scales as

$$\left\| \sum_{s \leq t} \epsilon_s \right\|_2 \propto \sigma\sqrt{t}.$$

Putting everything together, the distortion between h_t and h_t^* satisfies $\|h_t - h_t^*\|_2 \propto \sigma\sqrt{t}$ □

proposition 3 implies that with high probability, the trajectory of the generated hidden state h is confined within a cone centered at h^* whose radius grows at a rate $\propto \sigma\sqrt{t}$.

When mode collapse occurs at inference time, i.e., a generated sequence $x_{\leq t}$ collides with $y_{\leq t'}$, then their corresponding hidden states h and g must collide. Let $\|h^* - g^*\|_2$ be the minimum distance between h^* and g^*. By proposition 3, $\forall \varepsilon > 0$, $\exists c$,

$$P(\|h^* - g^*\|_2 > c \cdot \sigma\sqrt{t}) \leq \varepsilon$$

On the other hand, $\mathcal{L}_{\text{STP}}$ suppresses ϵ_t and consequently reduces σ, which decreases the lower bound of the probability of mode collapse.

G Implementation Details

If the training data already possesses a two-view structure, such as a $(query, answer)$ pair, one can leverage it by anchoring s at the beginning of the $query$ and t at the end of the $answer$. However, we suggest that r should be randomly selected to maximize the benefit of the STP loss. As demonstrated in our ablation study, fixing r at the end of the $query$ yields lower accuracy.

Typically, $h_t - h_s$ does not equal the hidden state of the isolated sub-sequence $x_{[s,t]}$. However, as discussed section C, we can view $h_t - h_s$ as the semantic evolution induced by the sub-sequence $x_{[s,t]}$ given the context $x_{\leq s}$. In this sense, $h_t - h_s$ acts as a context-aware hidden state, which is significantly more informative than the hidden state of $x_{[s,t]}$ computed in isolation. For example, given the prefix $\vec{v}_{\text{The}}, \vec{v}_{\text{capital}}, \vec{v}_{\text{of}}$, appending the token $\vec{v}_{\text{France}}$ shifts the overall meaning to $\vec{v}_{\text{Paris}}$. Conversely, given the prefix $\vec{v}_{\text{The}}, \vec{v}_{\text{language}}, \vec{v}_{\text{of}}$, appending $\vec{v}_{\text{France}}$ shifts the meaning to $\vec{v}_{\text{French}}$. Computing the hidden state of $\vec{v}_{\text{France}}$ separately loses this context and fails to capture the context-specific meaning of the tokens (see fig. 5).

We can also leverage $h_t - h_s$ to bypass unwanted tokens. For example, setting $s > 0$ allows us to skip the system prompt. Similarly, in multiple-choice Q&A, distractor choices that are semantically inconsistent with the $query$ are often located between the $query$ and the correct $answer$. In such cases, we can pick r and r' such that $x_{[s,r]}$ is the $query$ and $x_{[r',t]}$ is the correct $answer$, computing the STP loss as:

$$\mathcal{L}_{\text{STP}} = 1 - \cos(h_t - h'_r, h_r - h_s).$$

This formulation effectively skips the irrelevant choice branches in the middle.

Finally, the STP loss assumes that h_s, h_r, and h_t are collinear, which may not hold strictly in reality as geodesics can exhibit curvature. In practice, this implies that we must select a small λ to tolerate the angular deviation between $h_t - h_r$ and $h_r - h_s$. Indeed, our experiments consistently show that $\lambda \approx 0.01$ is effective across various models, datasets, and model sizes.

H Signal-to-Noise Ratio

Directly measuring Signal-to-Noise Ratio (SNR) in the latent representations of LLMs is intractable. In self-supervised learning, the decomposition of activations into "semantic signal" and "nuisance noise" is not explicitly observable without access to the ground-truth data manifold.

In this subsection, we formally show an information theoretic link between SNR and data efficiency and training accuracy. Hence we can validate our hypothesis via the predicted impact on them.

We model LLM training process as extracting information about a discrete target Y (tokens) from continuous latent representations X (hidden states). Let $Y \in \mathcal{V}$ be the discrete target token from a vocabulary of size $|\mathcal{V}|$. Let $X^m = \{X_i, 1 \leq i \leq m\}$ be a set of m hidden states that are conditionally i.i.d. given Y. The training objective is to minimize cross-entropy, which is asymptotically equivalent to minimizing the conditional entropy $H(Y|X^m)$.

Lemma 2 (Data Efficiency)**.**

$$H(Y|X^m) \geq H(Y) - m \cdot I(Y; X) \tag{5}$$

Proof. The goal is to show:

$$H(Y|X^m) \geq H(Y) - mI(X;Y)$$

By the definition of Mutual Information:

$$H(Y|X^m) = H(Y) - I(Y;X^m)$$

We need to bound $I(Y;X^m)$. Apply chain rule of mutual information,

$$I(Y;X^m) = H(X^m) - H(X^m|Y)$$

Since X_i are conditionally independent given Y:

$$H(X^m|Y) = \sum_{i=1}^{m} H(X_i|Y)$$

For the first term $H(X^m)$, by sub-additivity of entropy, the entropy of the joint distribution is always less than or equal to the sum of individual entropies (independence maximizes entropy):

$$H(X^m) \leq \sum_{i=1}^{m} H(X_i)$$

Substitute these back into the Mutual Information expansion:

$$I(Y;X^m) \leq \sum_{i=1}^{m} H(X_i) - \sum_{i=1}^{m} H(X_i|Y)$$

$$I(Y;X^m) \leq \sum_{i=1}^{m} (H(X_i) - H(X_i|Y))$$

$$I(Y;X^m) \leq \sum_{i=1}^{m} I(Y;X_i)$$

Since X_i are identically distributed, $I(Y;X_i)$ is the same for all i:

$$I(Y;X^m) \leq m \cdot I(Y;X)$$

Finally substitute this upper bound on Information back into step 1. Since we are subtracting a larger value, the result is a lower bound on entropy:

$$H(Y|X^m) = H(Y) - I(Y;X^m) \geq H(Y) - m \cdot I(Y;X)$$

□

Suppose $H(Y|X^m) \leq \epsilon$ after training, we have

$$\epsilon \geq H(Y|X^m) \geq H(Y) - m \cdot I(Y;X)$$

Recent theoretical work on infinite-width limits (Yang & Littwin, 2021) establishes that layer pre-activations converge to Gaussian distributions. Motivated by this, we model the local representation dynamics using a canonical Gaussian Channel approximation with additive noise. Specifically, we decompose $X = Z + N$, where Z is the latent signal, and $N \sim \mathcal{N}(0, \sigma^2 I)$ is the additive Gaussian noise. We define the Signal-to-Noise Ratio as

$$\mathrm{SNR} = \frac{\mathbb{E}[\|Z\|^2]}{\mathbb{E}[\|N\|^2]}$$

Under the Gaussian channel approximation, mutual information is a logarithmic function of SNR (Shannon, 1948):

$$I(X;Y) = \frac{1}{2}\log(1+\mathrm{SNR})$$

Substituting this capacity into lemma 2, we have

Corollary 2 (Signal-to-Noise Ratio)**.**

$$m \geq \frac{H(Y)-\epsilon}{\frac{1}{2}\log(1+\mathrm{SNR})} \tag{6}$$

Corollary 2 indicates that m is inversely proportional to $\log(1+\mathrm{SNR})$. Consequently, if the Semantic Tube works as expected, it will increase SNR and strictly lower the data requirement m.

Let $\hat{Y} = f(X^m)$ be the estimator of Y produced by the model. Let $P_e = P(\hat{Y} \neq Y)$ be the probability of error (incorrect token generation). Fano's Inequality (Cover & Thomas, 1991) provides a lower bound on the conditional entropy $H(Y|X^m)$ in terms of the error probability:

$$H(Y|X^m) \leq H_b(P_e) + P_e \log(|\mathcal{V}|-1)$$

where $H_b(P_e)$ is the binary entropy function. For LLMs, $|\mathcal{V}| \gg 1$, the term $P_e \log|\mathcal{V}|$ dominates $H_b(P_e)$. Hence we can simplify Fano's inequality to be:

$$H(Y|X^m) \leq P_e \log(|\mathcal{V}|-1) \tag{7}$$

Plug eq. (5) into eq. (7), immediate we get

Corollary 3 (Accuracy)**.**

$$P_e \gtrsim \frac{H(Y) - m \cdot \frac{1}{2}\log(1+\mathrm{SNR})}{\log|\mathcal{V}|} \tag{8}$$

Corollary 3 indicates that if we observe significant improvement on training accuracy, we know that SNR is higher.

I More Related Work

Our approach addresses the classic **Exposure Bias** problem (Bengio et al., 2015), originally identified in recurrent neural networks (RNNs) (Elman, 1990; Siegleman & Sontag, 1995). The problem arises because the model is trained with **Teacher Forcing** (Williams & Zipser, 1989)—conditioning on the ground-truth history—but must rely on its own potentially drifting predictions during inference. Although Maximum Likelihood Estimation ($\mathcal{L}_{\mathrm{NTP}}$ in the case of LLMs) is empirically effective, Huszár (2015) argues that it optimizes an objective different from generation quality, motivating our combined loss $\mathcal{L}_{\mathrm{NTP}} + \mathcal{L}_{\mathrm{STP}}$.

Our framework extends the philosophy of **Energy-Based Models** (EBMs) (LeCun et al., 2006), which learn to assign low energy to compatible configuration of variables. While EBMs and recent architectures like JEPA (LeCun, 2022) typically minimize energy at specific states, our approach invokes the Principle of Least Action to minimize the action—the integral of the Lagrangian along the generation trajectory. By enforcing geodesic constraints via $\mathcal{L}_{\mathrm{STP}}$, we generalize state-wise (or local) energy minimization to trajectory-wise action minimization, ensuring the generation follows the path of least resistance.

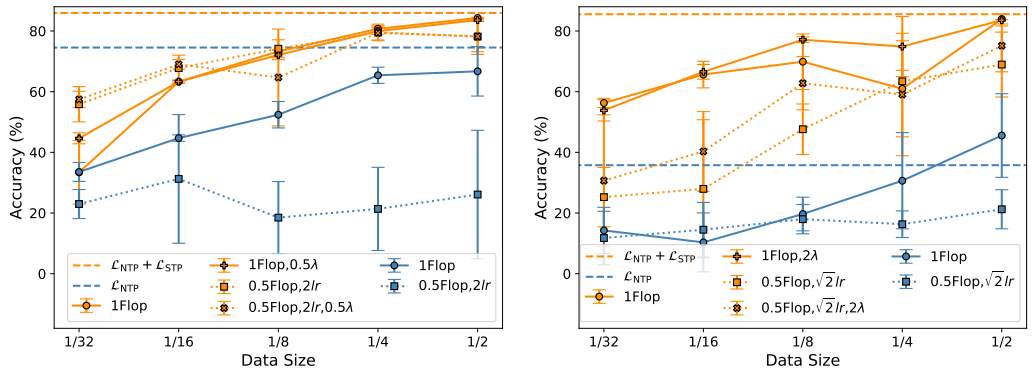
The Manifold Hypothesis (Kiani et al., 2024; Robinson et al., 2025; Whiteley et al., 2025) posits that learned representations form a simple and smooth manifold. Under the Geodesic Hypothesis, this structure is a natural consequence of the Principle of Least Action.

The Curvature Straightening Phenomenon (Hosseini & Fedorenko, 2023; Hénaff et al., 2021) observes that the training process tends to straighten the curvature between consecutive tokens. We interpret this as a manifestation of the underlying geodesic, which approximates a straight line.

The **Neural Tangent Kernel (NTK)** simplifies infinite-width dynamics (Jacot et al., 2018), a framework generalized to Transformers (Hron et al., 2020; Yang & Littwin, 2021) and compatible feature learning regimes (Yang & Hu, 2021). While Seleznova & Kutyniok (2022) note the importance of the depth-to-width ratio, modern LLMs typically operate in the requisite width $\gg$ depth regime.

J Data Efficiency

In this section we present the results of experiments on Llama3 3B and 8B using $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$, and $\frac{1}{32}$ of the dataset in fig. 8, where we see similar trend as in Llama3 1B (fig. 1).



(a) Llama3 3B (b) Llama3 8B

Figure 8: Semantic Tube (our approach) and regular fine-tuning with $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$, and $\frac{1}{32}$ dataset on (a) Llama3 3B and (b) Llama3 8B.

K Regular Expression Samples

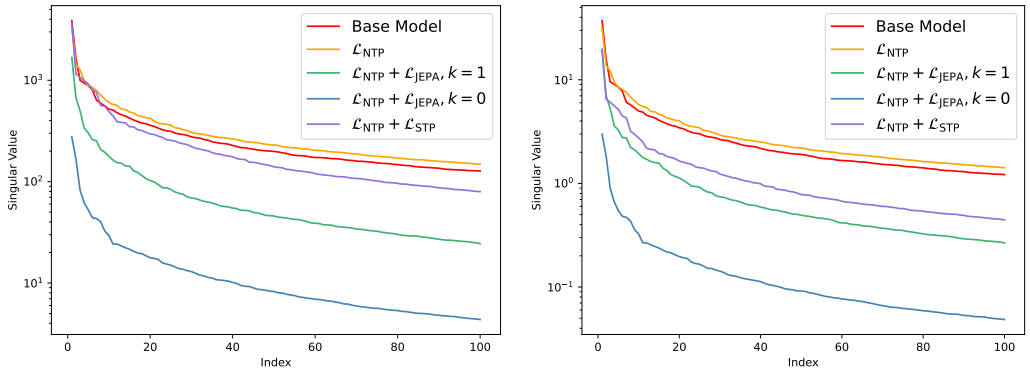
We list in table 2 a few samples from the SYNTH dataset that end with either "$.*$" or "$.*.*$", which are functionally equivalent.

Table 2: Regular expression samples from the SYNTH dataset that end with either "$.*$" or "$.*.*$", which are functionally equivalent.

Regular Expressions
.*([a-z])|([AEIOUaeiou])|([A-Za-z]).*
.*([A-Za-z]).*([0-9]).*.*
((dog)(.*)).*([AEIOUaeiou]).*
(dog).*((truck)|([A-Z])|([0-9])).*
.*(.)&([0-9])&(dog).*
.*(dog).*((.)*).*.*
.*dog.*[a-z].*.*

L SVD POLYMORPHISM

Following LLM-JEPA, we compute the singular value decomposition (SVD) of $\mathrm{Enc}(\mathrm{Text}) - \mathrm{Enc}(\mathrm{Code})$ to gain insight into the learned representations. Interestingly, we find (fig. 9) that Semantic Tube exhibits polymorphism: when the difference vectors $\mathrm{Enc}(\mathrm{Text}) - \mathrm{Enc}(\mathrm{Code})$ are normalized, the singular value spectrum aligns with LLM-JEPA; however, without normalization, it closely resembles regular fine-tuning. This indicates that Semantic Tube enforces structure on the directions (normalized vectors) while tolerating complexity on the raw vectors. We conjecture that this mechanism allows Semantic Tube to maintain flexibility and preserve diversity.



(a) Without Normalization

(b) With Normalization

Figure 9: SVD decomposition demonstrating Semantic Tube's polymorphism. **(a)** Without normalization, the SVD profile closely resembles regular fine-tuning. **(b)** With normalization, the SVD aligns with LLM-JEPA. Collectively, this indicates that Semantic Tube enforces a simple structure on the directions (normalized vectors) mapping Text to Code, while tolerating complexity in the unnormalized vectors. Note that the relative relationships among the base model, regular fine-tuning, and LLM-JEPA remain unchanged with or without normalization.

M TUNING λ

Semantic Tube introduces a single hyperparameter, λ. Empirically, we observe that the accuracy vs. λ curve is concave (fig. 4), typically peaking between 0.01 and 0.08 (table 3). Notably, this behavior persists across other variations: the accuracy curves remain concave, and the optimal λ consistently falls within the 0.01–0.08 range (see fig. 10). This validates (P4).

Table 3: Optimal λ values yielding maximum accuracy.

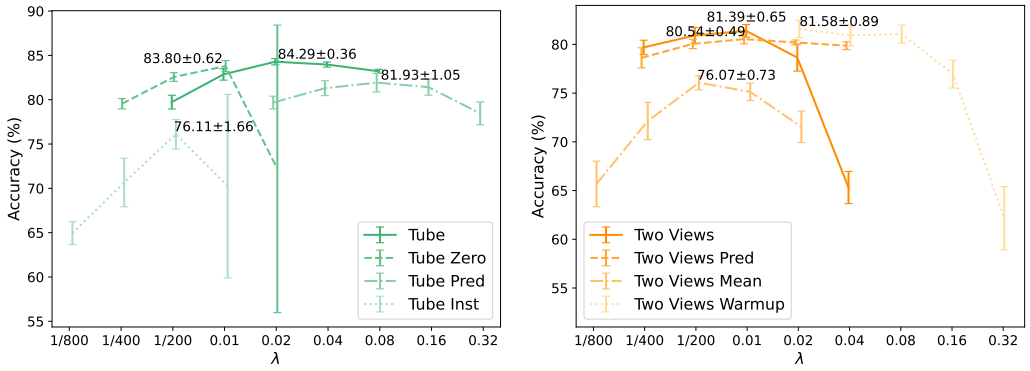
SYNTH	TURK	GSM8K	Spider	NQ	HS
0.02	0.04	0.005	0.04	0.16	0.02

Gemma2	Qwen3	R1 Dist	OLMo	OpenELM
0.005	0.02	0.04	0.01	0.04

Llama3 3B	Llama3 8B
0.01	0.0025

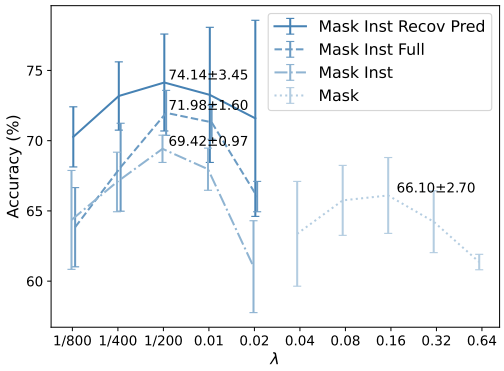
N ABLATION

We conducted extensive ablation studies on design decisions, establishing that $\mathcal{L}_{\text{STP}}$ yields superior performance compared to all variations (fig. 11). We specifically note that the **Pred** variant—which trains a linear projector P to minimize $\mathcal{L}_{\text{STP}} = 1 - \cos(P(h_r - h_s), h_t - h_r)$—results in degraded performance in all configurations. This validates (P5).



(a) Semantic Tube

(b) Two Views



(c) Mask

Figure 10: Tuning λ for various configurations of (a) Semantic Tube, (b) Two Views, and (c) Mask. In all cases, the accuracy vs. λ curve is concave. We also observe that when λ exceeds the optimal value, accuracy declines rapidly while the standard deviation increases sharply, indicating that $\lambda \ll 1$ is preferred.

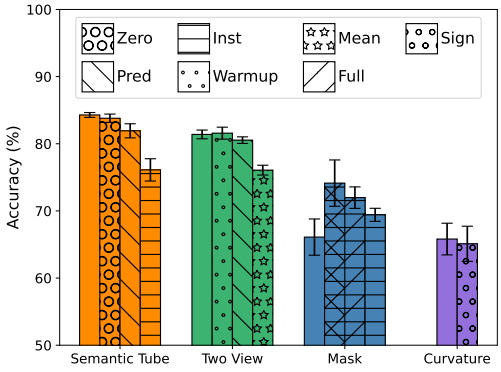


Figure 11: Ablation study. Semantic Tube (our approach) outperforms all variations. Within the Semantic Tube family, alternative configurations consistently degrade performance.

Semantic Tube: We ablate several variations of the Semantic Tube configuration:

- **Zero**: Instead of randomly picking s, this variation fixes the start index $s = 0$. The loss becomes $\mathcal{L}_{\text{STP}} = 1 - \cos(h_r - h_0, h_t - h_r)$.

- **Pred**: We introduce a learnable linear projector P and modify the loss to $\mathcal{L}_{\text{STP}} = 1 - \cos(P(h_r - h_s), h_t - h_s)$. aligns the approach more closely with the JEPA style, utlizing a non-identity predictor. P is randomly initialized and trained during fine-tuning.
- **Inst**: We incorporate instructions into the token sequence $x_{\leq t}$. These instructions consist of system prompt such as "`Convert natural language to regular expression`".

Two Views: This configuration adopts the LLM-JEPA style two-view structure, where *query* and *answer* represent two views of the same concept. Note that we retain the $\mathcal{L}_{\text{STP}}$ formulation but fix $s = 0$ and set r to the index of the last token of the *query*.

- **Warmup**: We linearly warm up λ throughout the training process.
- **Pred**: Identical to the Pred variation in the Semantic Tube configuration.
- **Mean**: Instead of the difference vector $h_r - h_s$, we use the average embedding $\frac{1}{r-s+1}\sum_{s\leq i\leq r} h_i$. Consequently, the loss becomes $\mathcal{L}_{\text{STP}} = 1 - \cos(\frac{1}{r-s+1}\sum_{s\leq i\leq r} h_i, \frac{1}{t-r+1}\sum_{r\leq j\leq t} h_j)$. This is inspired by BERT Mean Pooling (Kim et al., 2021).

Mask: This variation is inspired by BERT mask-and-recover training objective (Devlin et al., 2019). Given a token sequence $x_{\leq t}$, we randomly pick a span $[s, r]$ and replace the tokens within this span with the `[MASK]` token. Let $y_{\leq t}$ denote the masked sequence and $g_t = f(y_{\leq t})$. The loss is defined as $\mathcal{L}_{\text{mask}} = 1 - \cos(h_r - h_s, g_t)$. This can be interpreted as recovering the information of the masked tokens using the representation of the masked sequence $y_{\leq t}$.

- **Full**: Instead of aiming to match $h_r - h_s$, we target h_t. The loss becomes $\mathcal{L}_{\text{Mask}} = 1 - \cos(h_t, g_t)$, corresponding to the recovery of the full masked sequence rather than just the masked span.
- **Pred**: Identical to the Pred variation in the Semantic Tube configuration.
- **Inst**: Identical to the Inst variation in the Semantic Tube configuration.

Curvature: This variation is inspired by the curvature straightening objective (Hénaff et al., 2021). Let θ_i be the angle between $h_i - h_{i-1}$ and $h_{i+1} - h_i$. The loss is defined as $\mathcal{L}_{\text{Curvature}} = \frac{1}{t}\sum_{i\leq t} |\theta_i|$.

- **Sign**: Replaces $|\theta_i|$ with θ_i (allowing for signed curvature).

The fact that Pred yields inferior performance in both the Semantic Tube and Two Views configurations supports (P5).

The p-values comparing variations and options are presented in tables 4 and 5.

Table 4: Pairwise p-values comparing variation families. A cell is populated only if the mean accuracy of the row method exceeds that of the column method. p-values are computed using a paired, one-tailed t-test, restricted to the best-performing variant from each family.

	Two View	**Mask**	**Curvature**
LLM-JEPA2	1.14e-3	1.77e-3	3.04e-5
Two View		4.76e-3	5.10e-5
Mask			1.28e-4

Table 5: Pairwise p-values comparing options within each variation family. A cell is populated only if the mean accuracy of the row option exceeds that of the column option. p-values are computed using a paired, one-tailed t-test. Values exceeding 0.05 are struck through.

	Zero	**Pred**	**Inst**		**2View**	**Pred**	**Mean**
LLM-JEPA2	~~0.0534~~	2.03e-3	2.34e-4	**2View+Warmup**	~~0.265~~	0.0426	9.16e-6
+Zero		0.0185	5.56e-4	**2View**		~~0.0689~~	1.19e-6
+Pred			2.97e-4	**+Pred**			2.78e-4

	Inst,Recov	**Inst**	**Mask**		**Signed**
Mask+*all*	~~0.0629~~	0.0159	4.02e-3		
-Pred		2.34e-3	1.18e-3	**Curvature**	0.0368
-Recov,Pred			0.0103		