TOWARDS THE SCALING LAW OF NEURAL MODEL PROBES

Anonymous authors
Paper under double-blind review

ABSTRACT

Probing, the practice of training a lightweight classifier on frozen model representations and reading off its accuracy, has become a popular tool for studying what neural network models encode. Yet probes themselves have received little systematic study: design choices such as architecture, hidden width, and initialization are often made arbitrarily, and probing accuracy alone cannot separate information present in the representation from what the probe learns on its own. In this paper, we study the science of probes, and ask whether their behaviors follow scaling laws analogous to those of model pretraining. We use minimum description length (MDL; Voita & Titov, 2020) as our measure of probe faithfulness, and study its relationship with complexity of MLP probes (i.e., hidden width) and their training data size across data modalities, analyzed models, tasks, layers, and initialization schemes. We find that (1) faithfulness is non-monotone in probe width, exhibiting a characteristic U-shape with an optimum typically between widths of 64 and 256; (2) this optimum can be predicted from simple representation-level features, namely the layer depth and the accuracy of a linear probe; (3) dataset size affects faithfulness monotonically and exponentially; and (4) the scale of probe initialization shifts the optimum in a log-linear fashion, while maximal update parametrization flattens the faithfulness–complexity curve. Together, these findings set the first steps toward establishing a scaling law of probing and offer practical guidance for designing and configuring probes.

1 INTRODUCTION

Probing has become a widely used toolkit for examining what the internals of neural network models represent (Belinkov, 2022). Since its introduction by Ettinger et al. (2016), the practice of training a lightweight classifier (i.e., a *probe*) on frozen representations and reading off its accuracy has been applied to assessing many modern deep learning systems: from finding linguistic properties encoded in language models (Hewitt & Liang, 2019; Jawahar et al., 2019; Tenney et al., 2019) to understanding the representational space of pretrained visual (Alain & Bengio, 2017; Salin et al., 2022; Lysnæs-Larsen et al., 2025), audio (Pasad et al., 2021), and world (Li et al., 2023) models. With the development of large language models (LLMs) in recent years, probing has also gained attention as a core method for interpreting LLMs (Park et al., 2023; Gurnee et al., 2023, *inter alia*).

Yet the probes themselves, the fundamental backbone this progress depends on, have received little systematic study. Almost every study makes a series of design choices on factors such as probe architecture, hidden width, initialization strategy, optimizer with little guarantee in their optimality and comparability across settings. When tuning is performed, it is typically guided by validation accuracy and a heuristic and, to some extent, unsubstantiated belief that "simple enough" probes are faithful while "too complex" probes are not (Belinkov, 2022). This is unsatisfying for a measurement on which so much downstream interpretation depends. Intuitively, a faithful probe should be one that achieves high performance while relying minimally on its own capacity (Voita & Titov, 2020), and this is fundamentally a statement about the joint behavior of probe complexity and representation quality, not about either in isolation.

In this paper, we take the probe itself as our object of study. Rather than using probes to characterize a pretrained model, we ask: *how does a probe's behavior depend on its own design, and do scaling laws like those of model pretraining (Kaplan et al., 2020; Hoffmann et al., 2022; Cherti et al., 2023)*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

exist for probes? We use minimum description length (MDL; Voita & Titov, 2020) as our measure of probe faithfulness, where it views a probe as a code that transmits labels given representations, and decomposes its description length into two terms: the cost of transmitting the model, and the cost of transmitting the data given the model. A probe that faithfully reflects information in the representation pays a small model-transmission cost, while one that fits the task on its own pays a large one. MDL therefore measures both the quality of a probe and the cost of achieving that quality.

Using MDL, we conduct a systematic scaling study of MLP probes across modalities (text, vision, audio), models (Pythia, Llama-3.1, Qwen-3, CLIP, DINOv2, RayZer, Whisper), tasks, layers, and initialization schemes. Our findings form a coherent, and in places, surprising picture on the relationship between a probe's faithfulness and its complexity (e.g., hidden state width, training data size), as summarized below:

- **Faithfulness exhibits non-monotonic relationship to probe complexity.** Across settings, the MDL of a probe typically presents a characteristic U-shape as its hidden width increases, with an optimum usually occuring between widths of 64 and 256. A direct decomposition of the MDL curve shows that this U-shape is the result of interaction between two terms: data codelength falls rapidly and then plateaus, while model codelength rises throughout, and their crossover defines the optimum. In a few other settings, the MDL continues to decrease before flattening into a plateau rather than rising again, indicating that unbounded probe complexity is rarely useful. This suggests that excessive complexity yields diminishing—and often negative—returns to faithfulness.
- **Faithfulness monotonically increases when dataset size grows.** Across settings, the NLL ($\frac{\text{MDL}}{\text{dataset size}}$) of a probe presents an exponential relationship as dataset size increases. The NLL is predictable from dataset size and model width, similar to neural scaling law.
- **The optimum is predictable from cheap, representation-level features.** Across tasks and models, two simple features of the representation, the layer depth and the accuracy of a linear probe, are sufficient to predict the optimal hidden width with substantial R^2, suggesting an encouraging step towards a closed-form scaling law that practitioners can apply without sweeping over widths.
- **Initialization reshapes the scaling law.** The scale of probe weight initialization, parameterized by γ in $\boldsymbol{W}_{ij} \sim \mathcal{N}(0, d^{-2\gamma})$ (Yao et al., 2025), shifts the optimum in a log-linear fashion: smaller initialization ranges favor larger probes, and a single deterministic γ is, for many widths, demonstrably suboptimal. Under maximal update parametrization (μP; Yang et al., 2022), faithfulness changes from a U-shape to a monotonic decrease with diminishing returns, and the optimal learning rate becomes width-invariant, recovering in the probing setting the kind of zero-shot hyperparameter transfer that μP was designed to deliver in pretraining.

Taken together, these results suggest that probes exhibit lawful and predictable scaling behavior of their own. We view this work as a step toward a *science of probes*: a body of empirical regularities and design principles that gives the probing methodology the same kind of basis that scaling laws have given to model and data design, so that empirical findings drawn from probes can rest on a probe whose behavior is itself well understood.

2 Related Work

Probing in Deep Neural Network Models Probing involves training a simple classifier to predict some property from a model's representations (Alain & Bengio, 2017; Belinkov, 2022). It has been applied extensively to study linguistic and factual knowledge within language models, providing insights into how knowledge representations evolve across layers for different tasks (Conneau et al., 2018; Tenney et al., 2019). Also, given the recent advances in massive self-supervised or weakly-supervised vision and audio models, these models are now commonly used as frozen backbone encoders (Radford et al., 2021; Oquab et al., 2024; Radford et al., 2023). Understanding their learned representations is crucial for improving their design, and probing serves as an effective tool for this purpose (Salin et al., 2022; Pasad et al., 2021; Lysnæs-Larsen et al., 2025, *inter alia*).

Despite probing's widespread adoption, interpreting the results of probing experiments remains challenging, as probe performance does not always faithfully reflect how information is encoded by the underlying model (Hewitt & Liang, 2019; Ravichander et al., 2021). This fundamental challenge motivates the development of more principled approaches to the evaluation of probes.

Evaluation of Probes Accuracy, the common metric for probe quality (Cao et al., 2021), cannot distinguish information encoded in the model's representations from information memorized by the probe. Hewitt & Liang (2019) show that probes trained on random label assignments (control tasks) achieve accuracy comparable to those trained on genuine linguistic categories. Prior work has also turned to Pareto frontiers that balance accuracy and complexity to evaluate probes, yet this approach is uninformative under simple tasks and small probes (Pimentel et al., 2020). Minimum description length (MDL) addresses this challenge by jointly measuring probe quality and the effort required to achieve it (Voita & Titov, 2020), providing the foundation for our faithfulness definition.

3 METHODOLOGY

Minimum Description Length (MDL) for Probing Formally, probing uses a dataset $\mathcal{D} = \{(s_i, y_i)\}_{i=1}^{n}$, where s_i is a raw input (e.g., a sentence, image, or audio clip) and $y_i \in \mathcal{Y}$ ($|\mathcal{Y}| = K$) is the label for the property to be probed. Given an L-layer pretrained model, let $\boldsymbol{x}_i^{\ell} \in \mathbb{R}^{d_{\text{in}}}$ denote the layer-ℓ representation of s_i, from which a probe $p_{\boldsymbol{\theta}}(y \mid \boldsymbol{x})$ is trained to predict y_i.

From an information-theoretic perspective, the goal of a probe is to transmit $y_{\{1:n\}}$ given $\boldsymbol{x}^{\ell}_{\{1:n\}}$ with minimal codelength (Voita & Titov, 2020). This entails two costs: data codelength (bits to describe the labels given the model) and model codelength (bits to describe the model itself). If model $p_{\boldsymbol{\theta}}$ is known to the receiver, the optimal data codelength follows the Shannon–Huffman bound:

$$L_{p_{\boldsymbol{\theta}}}(y_{1:n} \mid \boldsymbol{x}^{\ell}_{1:n}) = -\sum_{i=1}^{n} \log_2 p_{\boldsymbol{\theta}}(y_i \mid \boldsymbol{x}_i^{\ell}). \tag{1}$$

In practice, we adopt online coding (Voita & Titov, 2020), a method that encodes both model and data at the same time, to estimate MDL. We choose timesteps $1 = t_0 < t_1 < \cdots < t_S = n$ and encode data by blocks. The initial block $y_{1:t_1}$ is encoded with a uniform code, costing $t_1 \log_2 K$ bits. For each stage $i = 1, \ldots, S-1$, a model $p_{\boldsymbol{\theta}_i}(y \mid \boldsymbol{x}^{\ell})$ is trained on the first i blocks $\{\boldsymbol{x}^{\ell}_{1:t_i}, y_{1:t_i}\}$ and used to encode the next block $y_{t_i+1:t_{i+1}}$. The total online codelength is given by Eq. 2:

$$L_{\text{online}}(y_{1:n} \mid \boldsymbol{x}^{\ell}_{1:n}) = t_1 \log_2 K - \sum_{i=1}^{S-1} \log_2 p_{\boldsymbol{\theta}_i}\big(y_{t_i+1:t_{i+1}} \mid \boldsymbol{x}^{\ell}_{t_i+1:t_{i+1}}\big), \tag{2}$$

While L_{online} does not separate model and data cost explicitly, we decompose it post hoc. We interpret the cross-entropy loss of the model trained on all data as the data codelength (Eq. 1), as the model can be viewed as known. The difference $L_{\text{online}} - L_{p_{\boldsymbol{\theta}}}$ is interpreted as the model codelength.

Probes' Scaling Size In this paper, we focus on two-layer MLP probes, defined as Eq. 3:

$$f(\boldsymbol{x}) = \boldsymbol{W}_2^{\top} \cdot \phi(\boldsymbol{W}_1^{\top} \boldsymbol{x} + \boldsymbol{b}_1) + \boldsymbol{b}_2, \tag{3}$$

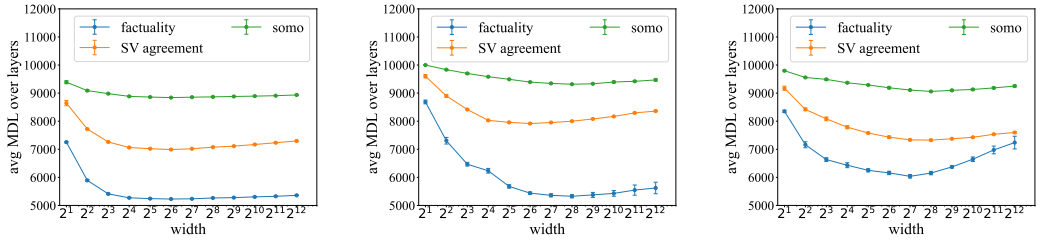
where $\boldsymbol{x} \in \mathbb{R}^{d_{\text{in}}}$, $\boldsymbol{W}_1 \in \mathbb{R}^{d_{\text{in}} \times w}$, $\boldsymbol{W}_2 \in \mathbb{R}^{w \times d_{\text{out}}}$, $\boldsymbol{b}_1 \mathbb{R}^{w}$, $\boldsymbol{b}_2 \in \mathbb{R}^{d_{\text{out}}}$, and $d_{\text{out}} = K$. d_{in} and d_{out} are the input and output dimensions; w is the hidden width that controls probe complexity. For each layer ℓ, we extract the layer's representations $\mathcal{X}_{\ell} = \{\boldsymbol{x}_i^{\ell}\}_{i=1}^{n}$ and train a series of probes with varying w, using Eq. 2 to compute the MDL of $\{y_{1:n}\}$ given $\mathcal{X}_{\ell}$. We define the **faithfulness** of a probe (at width w) as its MDL on $\mathcal{X}_{\ell}$, and take the average over all layers as a summary of the probe's overall faithfulness. By scaling w, we study the faithfulness–complexity relationship; by varying the training data size n, we study how faithfulness is influenced by training data.

We use MDL as our faithfulness metric because it is theoretically grounded, widely adopted (Perez et al., 2021; Du et al., 2023, *inter alia*), and better suited than alternatives. Prior work (Hewitt & Liang, 2019) also suggests a selectivity-based complexity sweet spot, though their control tasks do not directly transfer to our setting. More discussions about metric are in Appendix B.

4 FAITHFULNESS UNDER PROBE SCALING

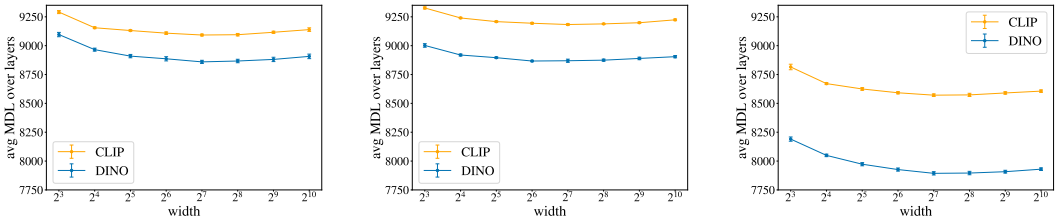
4.1 TEXT MODELS

Tasks To meaningfully evaluate probing behavior, the task should neither be trivial nor too difficult for the base model. We select three binary classification tasks: *subject–verb agreement* (SV



(a) Llama (b) Pythia (c) Qwen

Figure 1: Faithfulness–complexity relationship on three text models and three tasks.



(a) Car (b) Window (c) Flower

Figure 2: Faithfulness-complexity relationship on image classification.

agreement for short), *semantic odd man out* (SOMO; Conneau et al., 2018), and *factuality* (Rudinger et al., 2018). For all three tasks, we balance the positive and negative labels.

The *subject-verb agreement* task examines whether the subject and verb of an English sentence agree in number. The dataset is derived from the EWT and GUM Universal Dependencies treebanks (Nivre et al., 2020), using their native number annotations. Original sentences serve as positive examples, while negative samples are constructed by modifying the verb to disagree with its subject. The final dataset contains 15,830 training and 2,880 test sentences. [1]

In the *SOMO* task, a sentence is modified by replacing a random noun or verb with another noun or verb r, with the bigrams formed with r having frequencies comparable to those of the original bigrams. We use the dataset of Conneau et al. (2018), where the objective is to tell whether a sentence has been modified. The final dataset has 15k training samples and 10k testing samples.

The *factuality* task predicts whether an event, specified by a verb, actually happened. As many syntactic categories (e.g., negation, modal auxiliaries, determiners, adverbs) can convey information about the factuality of an event, this task requires integrating a broad range of semantic and syntactic cues. We use the UDS-IH2 dataset (Rudinger et al., 2018), yielding 16,436 training and 1,742 test samples after the filtering procedure described in Appendix C.1.2.

Results We use three state-of-the-art LLMs as base models: Pythia-6.9B (Biderman et al., 2023), Llama-3.1-8B (Llama Team, AI @ Meta, 2024), Qwen-3-8B (Yang, 2025). For *SV agreement* and *SOMO* tasks, $\boldsymbol{x}_i^\ell$ is the last-token representation of s_i at layer ℓ; for *factuality*, we concatenate the hidden representation of the final token in the sentence with that of the final token in the target verb span. In Fig. 1, we report the faithfulness–complexity curve for each (model, task) combination, averaged over 5 random seeds. **A clear U-shaped curve appears across all settings.** While the optimal width varies by task and model, they generally fall between 64 and 256. Notably, variance across seed is smaller near the optimum width, indicating that the probes are both faithful and stable.

4.2 Vision Models

Image Classification We use Open Images Dataset V7 (Kuznetsova et al., 2020) to classify whether objects from a given category are present in an image, treating the human-verified

[1] Testing set is not required for MDL computation, and is used only to verify reasonable probe performance.

confidence-1 and confidence-0 annotations provided in the dataset as positive and negative examples, respectively. We retain 10 categories with at least 10k positive and 10k negative samples and downsample them to balanced datasets. We use vlm-lens (Sheta et al., 2025) to extract hidden representations of these images from CLIP (Radford et al., 2021) and DINOv2 (Oquab et al., 2024), then repeat the probing experiments. Full results are in Fig. 12, Appendix C.2.3.

In Fig. 2, the U-shaped faithfulness–complexity curve persists across all object categories. Consistent with prior work (e.g., Hewitt & Liang, 2019), we observe that the trend of test set accuracy does not necessarily align with the trend of faithfulness; see Appendix C.2.2 for details.

Camera Parameter Regression We extend our exploration to a regression setting: probing camera extrinsic parameters from pre-trained RayZer (Jiang et al., 2025), a self-supervised multi-view 3D vision model. While RayZer's 8-layer transformer-based estimator internally predicts camera poses to enable photometric self-supervision, these predicted $SE(3)$ poses do not exactly align with the real-world pose space (Jiang et al., 2025), making the probing task non-trivial. We use DL3DV-10K dataset (Ling et al., 2024), which provides ground-truth camera extrinsic parameters. Because rotation parameters are bounded within $[-1, 1]$ while translation parameters are unbounded, we train separate probes for the two parameter sets.

While the definition of MDL does not naturally extend to continuous settings, the core part of online MDL is the negative log-likelihood $-\log p_{\boldsymbol{\theta}_i}(y_j \mid \boldsymbol{x}_j)$, the cost of encoding y_j at stage i. For regression, we instantiate this with a Gaussian distribution $p_{\boldsymbol{\theta}_i}(y_j \mid \boldsymbol{x}_j) = \mathcal{N}(f_{\boldsymbol{\theta}_i}(\boldsymbol{x}_j), \sigma^2)$, where $f_{\boldsymbol{\theta}_i}(\boldsymbol{x})$ is the probe's predicted value; when $\sigma = 1$ is fixed, the negative log-likelihood reduces to squared error. We therefore use a MDL-motivated proxy for faithfulness, calculated in a way similar to Eq. 4: encoding data by blocks, then summing the regression loss across different stages together. The initial constant term from Eq. 2 is omitted as it does not affect cross-layer comparisons. More details are provided in Appendix A.

$$\mathrm{MDL} = \sum_{i=1}^{S-1} \sum_{1 \le j \le t_i} \big(y_j - f_{\theta_i}(\boldsymbol{x}_j)\big)^2. \tag{4}$$

Fig. 13 in Appendix C.3 shows the faithfulness–complexity relationship for rotation and translation parameters. Unlike the classification settings, both curves decrease sharply before plateauing around width 32, suggesting a complexity threshold beyond which faithfulness gains are negligible.

4.3 Audio Models

We further extend our study to language identification tasks using Whisper (Radford et al., 2023), an encoder–decoder speech-to-transcript model. We probe Whisper's encoder representations using the Google FLEURS (Conneau et al., 2023) and Mozilla Common Voice (Ardila et al., 2020) datasets. The task requires the probe to predict the language of a given speech recording.

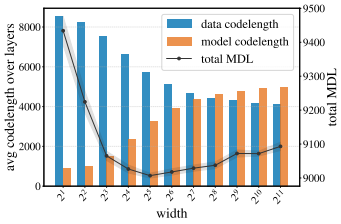
We first evaluate an in-distribution setting using five languages encountered during Whisper's pretraining: English, Spanish, French, Chinese, and Japanese (Fig. 14a, Appendix C.4.3). Subsequently, we construct five out-of-distribution binary classification tasks, each featuring a pair of closely related languages (Fig. 14b, Appendix C.4.3). In each pair, at least one language is unseen during Whisper's pretraining. Details of the language pairs are reported in Table 4 of Appendix C.4.

Similar to the regression setting (Fig. 13), Fig. 14a shows a monotonic decrease in MDL with diminishing returns as probe complexity increases. In Fig. 14b, MDL decreases before plateauing around width 256, while a few tasks exhibit a weak U-shape (e.g., lg-rw).

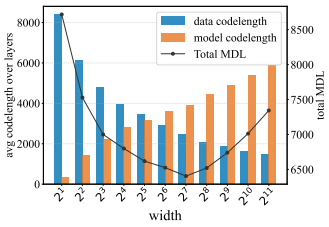
4.4 Interpreting the Scaling Behavior

Our experiments in § 4.1–4.3 reveal two recurring scaling patterns: in most settings, faithfulness exhibits a U-shaped curve with a clear optimum; in others, MDL decreases monotonically but with rapidly diminishing returns, plateauing beyond a complexity threshold. Both patterns suggest unlimited probe complexity is neither necessary nor beneficial for faithful probing.

To better understand the U-shaped behavior, we examine the two components of MDL: data codelength and model codelength. In Figure 3, we show the decomposition under two representative



(a) CLIP, car



(b) Qwen, factuality

Figure 3: Decompose MDL into data codelength (blue) and model codelength (orange) on image (left) and text (right) classification tasks. Total MDL (solid lines) correspond to the right y-axis, and data/model codelength (bars) correspond to the left y-axis.

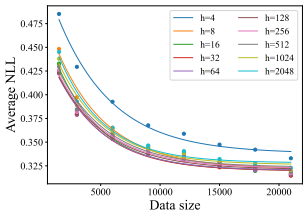


Figure 4: Faithfulness–probe's training data size relationship on Llama Factuality task.

settings that exhibit U-curve. As probe complexity increases, model codelength grows consistently, whereas data codelength decreases rapidly at first and then more gradually after the probe reaches a certain complexity threshold (e.g., width 512). In other words, in the low-complexity regime, the reduction in data codelength dominates, making the total MDL decrease. Beyond a complexity threshold, marginal gains in compression become negligible and the continued rise in model codelength takes over, generating the right-hand branch of the U-shaped curve. More results and discussions about MDL decomposition can be found in Appendix C.5.

5 Predicting the Optimum and Loss

5.1 Predicting the Optimum

In this section, we ask: can we **predict** this optimum from easily accessible representation-level properties, without training probes of every size?

Our goal is to predict the optimal width w^* (on a log scale) using intrinsic properties of the representation. As hidden representations could differ substantially across modalities, we fit separate models for text (3 base models, 3 tasks) and vision (2 base models, 10 object categories from Open Images). Unlike in § 4, where faithfulness is summarized by averaging MDL over layers, here we treat each layer as an independent observation, enabling finer-grained prediction. We consider five candidate predictors for w^*: (a) layer index ℓ, (b) mean and (c) variance of $\mathcal{X}_\ell$, (d) variance explained by the top-20 PCA components, and (e) the accuracy of a linear probe trained on the same task, as a proxy for representational separability.

We normalize all predictors to zero mean and unit variance, then apply a likelihood-ratio test to identify those that are genuinely predictive. We find that (a) layer index ℓ and (e) linear probe's accuracy are consistently significant across both modalities. Fitting a linear regression on these two predictors gives us the following scaling law (Eq. 5) for the optimal probe width:

$$\log_2(w^*) = b_\ell \cdot \ell + b_{\text{acc}} \cdot \text{acc} + b \tag{5}$$

Regression coefficients and R^2 for text and vision results are in Table 2. Specifically, layer index ℓ is the most dominant predictor: a univariate regression on ℓ alone achieves an R^2 of 0.462 for vision and 0.246 for text. While the signs of both coefficients differ across modalities, they are consistently significant predictors of the optimal width, suggesting they capture fundamental information about the representation regardless of modality. We exemplify a standard framework for principled probing: in large scale probing experiments, one could fit a scaling law on a small set of preliminary results, then use the resulting rule to guide further width selection.

5.2 Predicting the Loss with Data Size

We extend our study on the effect of training data size of probes. Fig. 4 shows a clear exponential relationship between faithfulness and data size. Results for more models and tasks are in Fig. 9. Moreover, as long as the dataset size is reasonably large (≥ 3000), the optimal hidden size is stable. More evidence can be found in Fig. 10 in Appendix C.1.4.

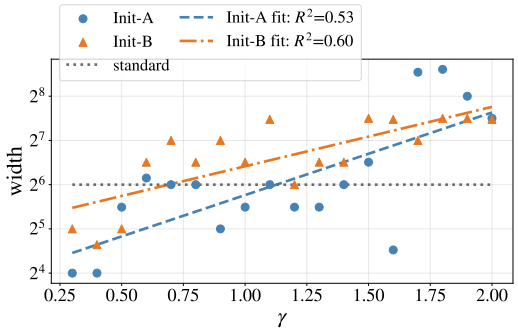


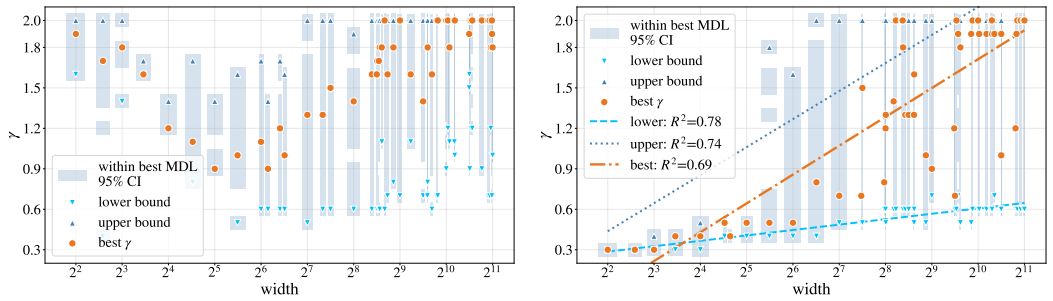
Figure 6: Optimal width (log scaled) under each γ under two initialization strategies. Lines are linear regression fit.

	b_l (95% CI)	b_{acc} (95% CI)	b (95% CI)	R^2
text	1.27 ($\pm$0.12)	$-$0.49 ($\pm$0.12)	5.93 ($\pm$0.12)	0.29
image	$-$1.78 ($\pm$0.26)	0.62 ($\pm$0.26)	7.28 ($\pm$0.15)	0.56

Table 2: Regression coefficients (with 95% two-sided confidence interval) and R^2 for Eq. 5.

Strategy	b_γ	b_{const}	R^2
Init-A	1.87 ± 0.93	3.90 ± 1.17	0.53
Init-B	1.34 ± 0.58	5.07 ± 0.74	0.60

Table 3: Regression coefficients (with 95% two-sided confidence interval) and R^2 for Eq. 8, under two initialization strategies.



(a) Init-A (b) Init-B

Figure 7: Optimal initialization scale γ and its near-optimal range across probe widths under Init-A and Init-B. Orange dots: the optimal γ (lowest mean MDL). Shaded regions: γ values whose mean MDL falls within the 95% confidence interval of the best MDL. Solid line: linear regression fitted to the optimal γ. Dashed and dotted lines: regressions fitted to the lower and upper boundaries of the near-optimal range, respectively.

We next ask: can we **predict** the loss when the specific model width w and data size is given? Following neural scaling laws (Kaplan et al., 2020; Hoffmann et al., 2022), we directly predict the average negative log likelihood (NLL; $\frac{\text{MDL}}{x_1}$) with the number of training samples (x_1) and log probe width (x_2) in Eq. 6 (an explanation for the design is in Appendix D.2):

$$\text{NLL} = (a_1 + d\,x_2)\exp(b_1 x_1) + a_2\exp(b_2 x_2) + a_3\exp(b_3 x_2) + c \quad (b_1, b_2 < 0,\ b_3, d > 0) \quad (6)$$

Across 9 settings (3 models $\times$ 3 tasks), 6 out of them achieve $R^2 \geq 0.9$. We show the regression results for Llama in Table 1. Full results are shown in Table 5 in Appendix D.1.

Task	a_1	b_1	d	a_2	b_2	a_3	b_3	c	R^2
Verb Agreement	0.191	-1.39×10^{-4}	0.001	0.254	-1.32	0.043	0.052	0.415	0.988
Factuality	0.156	-2.44×10^{-4}	$\leq 10^{-10}$	0.310	-1.04	0.003	0.171	0.308	0.950
SOMO	0.074	-1.67×10^{-4}	0.002	0.089	-0.651	0.005	0.097	0.592	0.978

Table 1: Fitted parameter values and R^2 for each task of Llama model.

6 Initialization Shapes the Faithfulness–Complexity Curve

Initialization is known to affect MLPs' (Jacot et al., 2018; Mei et al., 2018, *inter alia*) and LLMs' (Zhu et al., 2021; Trockman & Kolter, 2023; Zheng et al., 2026, *inter alia*) training dynamics. Motivated by this, we study two initialization strategies and show that each reshapes the faithfulness–complexity curve in a principled way. All experiments use Llama-3.1-8B on *SOMO*, our most challenging text-based task, whose optimal width under standard initialization is 2^6 (Fig. 1a).

6.1 Initialization Scale Shifts the Optimum

Initialization scale has been shown to affect LLMs' training dynamics and preference for reasoning versus memorization (Yao et al., 2025). We investigate whether it also affects the optimum width of probes. Following Yao et al. (2025), we initialize each parameter matrix via:

$$\boldsymbol{W}_{i,j} \sim \mathcal{N}\left(0,\ \left(d_1^{-\gamma}\right)^2\right), \tag{7}$$

where γ is the hyperparameter that controls initialization scale, and larger γ corresponds to smaller scale. Note that many default initialization methods, e.g., LeCun initialization (LeCun et al., 2002) and He initialization (He et al., 2015) correspond to $\gamma = 0.5$. For an MLP probe defined as Eq. 3, we study two initialization strategies:[2]

- **Init-A**: Both $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are initialized as Eq. 7.
- **Init-B**: $\boldsymbol{W}_1$ follows Eq. 7, while $\boldsymbol{W}_2 = 0$, inspired from LoRA initialization (Hu et al., 2021).

We vary γ from 0.3 to 2.0 in steps of 0.1, and evaluate 50 probe widths: $2^2, 2^{5/2}, 2^3, \cdots, 2^{21/2}, 2^{11}$, plus 31 randomly sampled widths within $[4, 2048]$. For each γ and width, we follow previous steps to measure the probe's faithfulness. Full results are in Appendix F.

Optimal complexity under different γ As shown in Fig. 6, increasing γ (smaller initialization) consistently shifts the optimal width toward larger values under both initialization strategies (the grey dotted line represents standard initialization). This relationship is approximately log-linear:

$$\log_2(w^*) = b_\gamma\, \gamma + b_{\text{const}}, \tag{8}$$

with regressed coefficients and R^2 reported in Tab. 3.

Optimal γ under different complexity Then, we fix the probe complexity and identify the initialization scale achieving the lowest MDL. To better quantify initialization sensitivity, we define the **near-optimal γ set** as the set of γ values whose MDL falls within the 95% confidence interval of the best γ, with formal definitions provided in Appendix F. Fig 7 shows that the two initialization strategies exhibit qualitatively different trends: under Init-B, the optimal γ increases monotonically with width, whereas under Init-A it first decreases and then rises, with a turning point near width 32. Under both strategies, the near-optimal range is substantially narrower for smaller probes, indicating greater sensitivity to initialization at low complexity.

Moreover, under Init-A, the standard initialization scale $\gamma = 0.5$ falls outside the near-optimal range for many widths, suggesting that the default initializations do not necessarily guarantee the most faithful probing. Under Init-B, both the optimal γ and near-optimal range increase with width, indicating a systematic interaction between initialization scale and probe complexity.

6.2 μP Reshapes the Curve and Transfers the Optimal Learning Rate

Under μP initialization (Yang et al., 2022), many optimal hyperparameters remain stable across neural network sizes, allowing for zero-shot hyperparameter transfer. It motivates two questions: (1) under μP initialization where "each layer is updated on the same order independent of width" (Yang et al., 2022), does the U-shaped faithfulness–complexity relationship persist? (2) when using faithfulness as the evaluation metric, does the optimal learning rate transfer across widths?

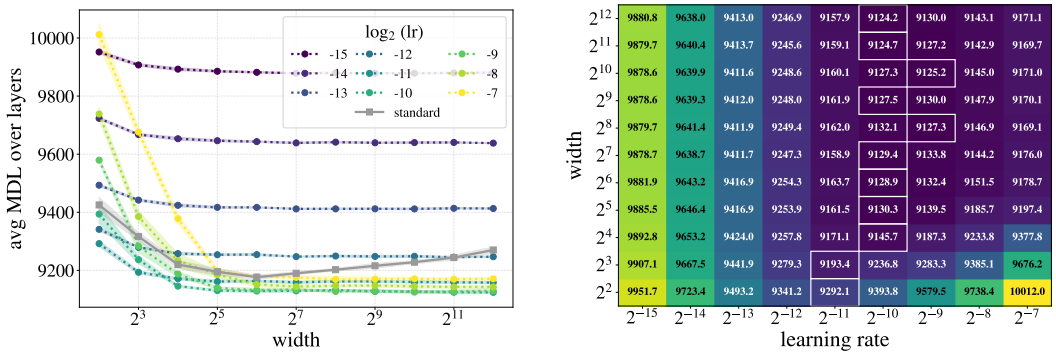
For an MLP defined in Eq. 3, its μP initialization is

$$\boldsymbol{W}_1 \sim \mathcal{N}(0, 1/d_{\text{in}}),\ \boldsymbol{W}_2 \sim \mathcal{N}(0, 1/w^2),\ b_{\{1,2\}} = 0,$$

with Adam (Kingma & Ba, 2017) learning rates $\eta_{\boldsymbol{W}_1} = \eta_{b_1} = \eta_{b_2} = \eta,\ \eta_{\boldsymbol{W}_2} = \eta w^{-1}$, where η is the base learning rate. For each width, we evaluate $\eta \in \{2^{-15}, \ldots, 2^{-7}\}$ and report faithfulness.

As shown in Fig. 8a, under μP, the U-shaped curve observed with standard initialization (grey line) is replaced by a monotonic decrease followed by a plateau. This indicates that the faithfulness–complexity relationship depends on the parametrization: standard initialization and μP exhibit distinct scaling behaviors. Moreover, μP initialization achieves comparable or even lower MDL (with

[2]Setting $\boldsymbol{W}_1$ to zero would cause no updating, as both forward and backward results are zero.



(a) Faithfulness-complexity relationship under different learning rates

(b) MDL across different widths and learning rates under μP initialization

Figure 8: Results under μP initialization.

learning rate range 2^{-7}–2^{-11}) than the optimum under standard initialization. This further suggests that standard initialization schemes do not necessarily yield the most faithful probes.

Fig. 8b further shows that the optimal base learning rate remains stable at $\eta^* = 2^{-10} \approx 0.97 \times 10^{-3}$ for all widths $w \geq 16$. The deviation at widths 4 and 8 is expected, since μP is an asymptotic theory that may not apply at small widths. Practically, these results suggest that under μP, one can tune the learning rate at a moderate probe width and reuse it across a broad range of larger widths.

7 Conclusion and Discussion

Our investigation into the scaling behavior of probes reveals that they exhibit principled and predictable scaling behaviors. We provide three primary contributions:

First, we demonstrate that the faithfulness of probe exhibits a non-monotonic relationship with its complexity (§ 4). Excessive complexity yields diminishing (plateaued MDL curves) or even negative (U-shaped curves) faithfulness returns. On the other hand, the effect of data size on faithfulness is exponential (§ 5.2).

Second, we exemplify a principled workflow for large-scale probing experiments (§ 5.1): practitioners can fit a customized scaling law with layer depth and linear-probe accuracy as predictors to guide complexity selection (Eq. 5), avoiding hyperparameter sweeping. Eq. 5 is practically useful for narrowing down a width choice. For a predicted optimal width $\hat{w}$, round$(\hat{w}) \pm 1$ captures the true optimal width w^* 53.4% of the time for text and 75.0% for vision, significantly outperforming the top-3 most frequent width based guessing baseline (45.0% and 49.2%, respectively). We also created an effective regression model for predicting NLL (§ 5.2), which is a counterpart for the neural scaling law.

Finally, we find that initialization shifts the faithfulness–complexity relationship (§ 6.1). Smaller initialization scales systematically shift the optimum towards higher complexity (Eq. 5); notably, the default initialization scale (0.5) often fails to achieve optimal faithfulness for a given width. Under μP, the faithfulness–complexity relationship follows a different scaling curve. μP can also achieve comparable or lower MDL and enables zero-shot learning-rate transfer across probe widths.

Future work could extend our faithfulness–complexity perspective to sparse autoencoders and transcoders, studying how latent width and sparsity shape the recovery of causally relevant features. Combining MDL-inspired evaluation with feature-level interventions could provide a more principled basis for width selection than reconstruction fidelity alone. Such work would connect a science of probes to mechanistic analysis, helping distinguish information encoded in the model from structure introduced by the interpretability method itself.

AI USE STATEMENT

The core method development in this research does not involve LLMs as any important, original, or non-standard components. We only use LLM for minor editing (e.g., grammar, spelling, word choice).

REPRODUCIBILITY STATEMENT

All experiments in this paper are conducted with fixed random seeds. We will release our codebase, datasets, and detailed configuration files containing all hyperparameters. With these seeds, datasets, and hyperparameters, our experiments are fully reproducible.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*, 2017.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. Common Voice: A Massively-Multilingual Speech Corpus. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Twelfth LREC*, pp. 4218–4222, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *CL*, 48(1):207–219, March 2022.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *ICML*, pp. 2397–2430. PMLR, 2023.

Steven Cao, Victor Sanh, and Alexander Rush. Low-complexity probing via finding subnetworks. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 NAACL: Human Language Technologies*, pp. 960–966, Online, June 2021.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *2023 CVPR*, pp. 2818–2829. IEEE, 2023. doi: 10.1109/cvpr52729.2023.00276. URL http://dx.doi.org/10.1109/CVPR52729.2023.00276.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In Iryna Gurevych and Yusuke Miyao (eds.), *ACL*, pp. 2126–2136, Melbourne, Australia, July 2018.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. FLEURS: FEW-Shot Learning Evaluation of Universal Representations of Speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pp. 798–805, January 2023. doi: 10.1109/SLT54892.2023.10023141.

Kevin Du, Lucas Torroba Hennigen, Niklas Stoehr, Alex Warstadt, and Ryan Cotterell. Generalizing backpropagation for gradient-based interpretability. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *ACL*, Toronto, Canada, July 2023. URL https://aclanthology.org/2023.acl-long.669/.

Allyson Ettinger, Ahmed Elgohary, and Philip Resnik. Probing for semantic evidence of composition by means of simple classification tasks. In *Proceedings of the 1st workshop on evaluating vector-space representations for nlp*, pp. 134–139, 2016.

Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. Finding neurons in a haystack: Case studies with sparse probing. *TMLR*, 2023. URL https://openreview.net/forum?id=JYs1R9IMJr.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, pp. 1026–1034, 2015.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *EMNLP-IJCNLP*, pp. 2733–2743, Hong Kong, China, November 2019.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models, 2022. URL https://arxiv.org/abs/2203.15556.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *NeurIPS*, volume 31. Curran Associates, Inc., 2018.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What does BERT learn about the structure of language? In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *ACL*, pp. 3651–3657, Florence, Italy, July 2019. doi: 10.18653/v1/P19-1356. URL https://aclanthology.org/P19-1356/.

Hanwen Jiang, Hao Tan, Peng Wang, Haian Jin, Yue Zhao, Sai Bi, Kai Zhang, Fujun Luan, Kalyan Sunkavalli, Qixing Huang, et al. RayZer: A self-supervised large view synthesis model. *arXiv preprint arXiv:2505.00702*, 2025.

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, 2020. URL https://arxiv.org/abs/2001.08361.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. URL https://arxiv.org/abs/1412.6980.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020.

Yann LeCun, Léon Bottou, Genevieve B Orr, and Klaus-Robert Müller. Efficient backprop. In *Neural networks: Tricks of the trade*, pp. 9–50. Springer, 2002.

Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *ICLR*, 2023. URL https://openreview.net/forum?id=DeG07_TcZvT.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *CVPR*, pp. 22160–22169, 2024.

Llama Team, AI @ Meta. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Jacob Lysnæs-Larsen, Marte Eggen, and Inga Strümke. Probing the probes: Methods and metrics for concept alignment, 2025. URL https://arxiv.org/abs/2511.04312.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 4034–4043, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Causal Representation Learning Workshop at NeurIPS*, 2023. URL https://openreview.net/forum?id=T0PoOJg8cK.

Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. Layer-Wise Analysis of a Self-Supervised Speech Representation Model. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. True few-shot learning with language models. In *NeurIPS*, 2021. URL https://openreview.net/forum?id=ShnM-rRh4T.

Tiago Pimentel, Naomi Saphra, Adina Williams, and Ryan Cotterell. Pareto probing: Trading off accuracy for complexity. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *EMNLP*, pp. 3138–3153, 2020. URL https://aclanthology.org/2020.emnlp-main.254/.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *ICML*, 2023.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the probing paradigm: Does probing accuracy entail task relevance? In *EACL*, 2021. URL https://aclanthology.org/2021.eacl-main.295/.

Rachel Rudinger, Aaron Steven White, and Benjamin Van Durme. Neural models of factuality. In *NAACL*, pp. 731–744, June 2018. URL https://aclanthology.org/N18-1067/.

Emmanuelle Salin, Badreddine Farah, Stéphane Ayache, and Benoit Favre. Are Vision-Language Transformers Learning Multimodal Representations? A Probing Perspective. *AAAI*, 36(10): 11248–11257, June 2022.

Hala Sheta, Eric Huang, Shuyu Wu, Ilia Alenabi, Jiajun Hong, Ryker Lin, Ruoxi Ning, Daniel Wei, Jialin Yang, Jiawei Zhou, Ziqiao Ma, and Freda Shi. From behavioral performance to internal competence: Interpreting vision-language models with vlm-lens. In *EMNLP (System Demonstration)*, 2025.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *ACL*, pp. 4593–4601, July 2019. URL https://aclanthology.org/P19-1452/.

Asher Trockman and J. Zico Kolter. Mimetic initialization of self-attention layers. In *ICML*, 2023.

Elena Voita and Ivan Titov. Information-theoretic probing with minimum description length. In *EMNLP*, pp. 183–196, November 2020.

An Yang. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Greg Yang, Edward J. Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer, 2022. URL https://arxiv.org/abs/2203.03466.

Junjie Yao, Zhongwang Zhang, and Zhi-Qin John Xu. An analysis for reasoning bias of language models with small initialization. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=4HQaMUYWAT.

Jianqiao Zheng, Xueqian Li, Hemanth Saratchandran, and Simon Lucey. Structured initialization for vision transformers. In *NeurIPS*, 2026.

Chen Zhu, Renkun Ni, Zheng Xu, Kezhi Kong, W Ronny Huang, and Tom Goldstein. Gradinit: Learning to initialize neural networks for stable and efficient training. In *NeurIPS*, 2021.

A Justification for Extending MDL to Continuous Setting

MDL's online coding interpretation is: the cost of encoding label y_i given $\boldsymbol{x}_i$ under model $\boldsymbol{\theta}$ is $-\log p(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})$. In Eq. 2, online MDL is defined through the negative log-likelihood accumulated across coding stages, where the first term $(t_1 \log_2 K)$ is constant under the same task and dataset. The core part is

$$\sum_{i=1}^{S-1} \log_2 p_{\boldsymbol{\theta}_i}\left(y_{t_i+1:t_{i+1}} \mid \boldsymbol{x}^{\ell}_{t_i+1:t_{i+1}}\right)$$

For classification, $p_{\boldsymbol{\theta}}$ is a categorical distribution parameterized by softmax, and $-\log p$ is exactly cross-entropy. For regression, we need to choose a distribution over continuous outputs. The natural choice is Gaussian:

$$p_{\boldsymbol{\theta}}(y \mid \boldsymbol{x}) = \mathcal{N}(f_{\boldsymbol{\theta}}(\boldsymbol{x}),\ \sigma^2),$$

where $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ is the probe's predicted value. Then:

$$-\log p_{\boldsymbol{\theta}}(y \mid \boldsymbol{x}) = \frac{(y - f_{\boldsymbol{\theta}}(\boldsymbol{x}))^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2)$$

This is exactly square loss up to a constant and scaling. Assume that $\sigma = 1$ fixed across all widths and stages, we could use square loss as a proxy for online MDL calculation.

While this extension is principled, we note two minor caveats regarding its application in our camera parameter regression task:

1. **Manifold Geometry**: Camera rotation parameters reside on $SO(3)$ rather than $\mathbb{R}$. Consequently, the Gaussian distribution serves as a local approximation of the true underlying geometry.
2. **Variance Assumption**: This proxy assumes that σ is constant across different probe widths and coding stages.

B Discussions about the selection of faithfulness metric

A faithful probe should "achieve high performance while relying minimally on its own capacity" (Voita & Titov, 2020), so any valid measure must reflect **both** parts: the quality of the probe's performance, and the cost to achieve it.

From an information-theoretic perspective, MDL is by far the most suitable metric for probe faithfulness:

$$\mathcal{L}_{\text{MDL}}(\text{model}, \text{data}) = \mathcal{L}(\text{model}) + \mathcal{L}(\text{data} \mid \text{model})$$

. The data codelength term $\mathcal{L}(\text{data} \mid \text{model})$ measures how well the model encodes data. The model codelength term $\mathcal{L}(\text{model})$ measures how much information is required to specify the probe itself. Existing literature also shows that MDL is highly reliable, robust to hyperparameters compared to accuracy, and widely adopted in interpretability (Perez et al., 2021; Du et al., 2023, *inter alia*).

We also surveyed a number of alternative candidates, but none of them satisfy both parts in our setting:

1. Raw accuracy/loss. As discussed by (Hewitt & Liang, 2019), a high accuracy or low loss could simply because the probe learns the task on its own.
2. Selectivity (Hewitt & Liang, 2019). In principle, selectivity satisfies both parts: real-task accuracy is the quality term, and control-task accuracy shows how much does a probe memorizes this task. Results in (Hewitt & Liang, 2019) (Figure 4, Rank/Hidden Dim, MLP-1) show that selectivity rises and then falls as probe hidden size increases, on both POS tagging and dependency edge prediction tasks. **We view it as an evidence of the existence of sweet spot in complexity.** However, their control tasks are designed for sequence tagging, and adapting them to common tasks like sentence classification remains difficult at present.

Given these considerations, MDL is the most rigorous and applicable metric for this study.

C Detailed Settings and Results of § 4 and § 5

C.1 Text Models

C.1.1 Hyperparameters

Text-based probes are trained with a 10^{-3} learning rate and 256 batch size, for up to 10 epochs with early stopping.

C.1.2 Sample Selection Criteria for Factuality Task

In UDS-IH2 dataset (Rudinger et al., 2018), each sample is annotated independently by two annotators, with both a binary factuality label for the verb's factuality and a confidence score ranging from 1 to 4. When the two annotations disagree, we retain the label with confidence 4 if the other annotation has confidence at most 2.

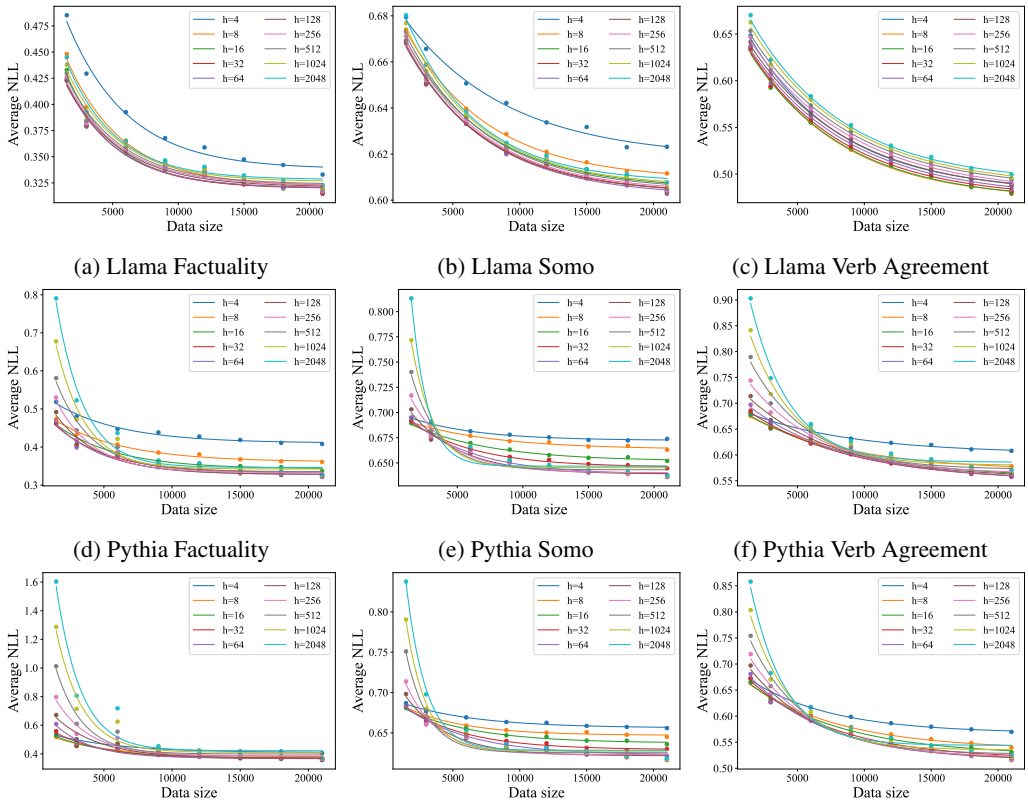
Note that "events" are specified by verbs, as a sentence can have multiple verbs, whose labels may vary. For example, in the following sentence *"'Our new lunar transportation system utilizes a unique architecture that will establish the equivalent of a two-way highway between the Earth and the Moon,' Kistler told SPACE.com"*, verb "established" did not happen, while "tell" happened.

C.1.3 Full Result for Dataset Size Scaling

Full results for data size scaling are shown in Figure 9.

C.1.4 Optimal Width for Different Dataset Size

The full result for data size scaling are shown in Figure 10. We can draw the conclusion from these heatmap, that while dataset size is important in probing performance (as is shown in Fig. 9), as long as the dataset size is reasonably large (≥ 3000), the optimal hidden size is stable. Thus, one can use



(a) Llama Factuality (b) Llama Somo (c) Llama Verb Agreement

(d) Pythia Factuality (e) Pythia Somo (f) Pythia Verb Agreement

(g) Qwen Factuality (h) Qwen Somo (i) Qwen Verb Agreement

Figure 9: Faithfulness–data size relationship on classification tasks in text models.

a small data size to determine the minima of U shape curve and then use larger dataset to get best probing performance.

C.2 Vision Models

C.2.1 Hyperparameters

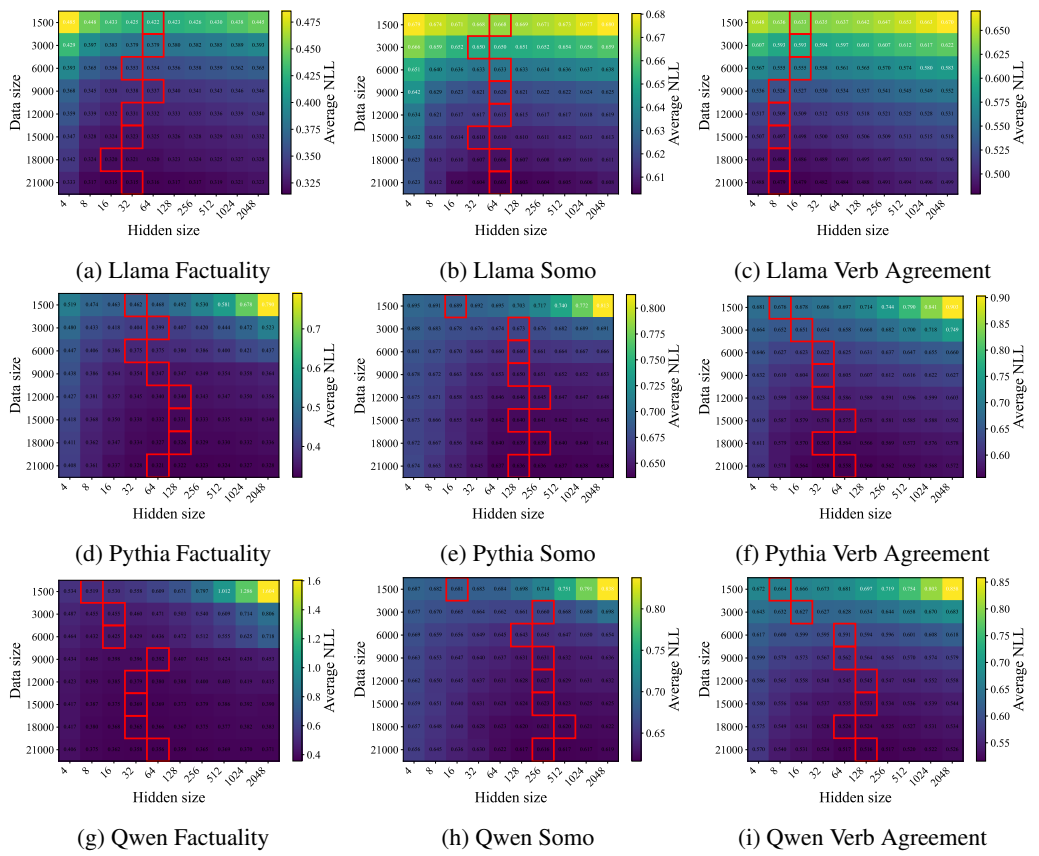
Text-based probes are trained with a 10^{-3} learning rate and 128 batch size, for up to 10 epochs with early stopping.

C.2.2 Accuracy May not Match with MDL in Trend

We observe that the trend of accuracy (on testing set) does not necessarily align with the trend of faithfulness. For example, in *flower*-presence classification, the width achieving the lowest MDL (2^7) on CLIP (Fig. 2c) counterintuitively yields the lowest average accuracy (Fig. 11c). Similarly, MDL on CLIP exhibits a turning point around width 2^6 (Fig. 2b), whereas the accuracy continues to decrease monotonically (Fig. 11b). These observations are consistent and again confirm that accuracy may fail to reflect the faithfulness of probing.

C.2.3 Image Classification Full Results

Full results on image classification tasks on the remaining categories are shown in Fig. 12.

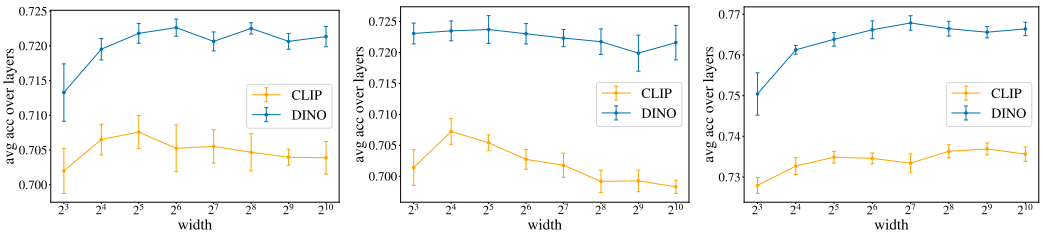


(a) Llama Factuality (b) Llama Somo (c) Llama Verb Agreement

(d) Pythia Factuality (e) Pythia Somo (f) Pythia Verb Agreement

(g) Qwen Factuality (h) Qwen Somo (i) Qwen Verb Agreement

Figure 10: Optimal width heatmap on classification tasks in text models.

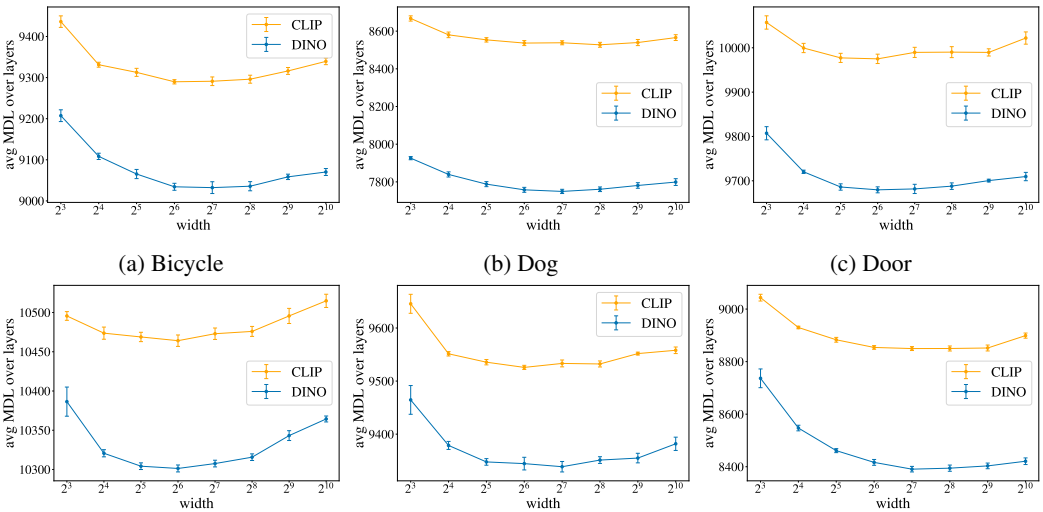


(a) Car, acc (b) Window, acc (c) Flower, acc

Figure 11: Image classification accuracy on Openimages, which may not show exactly the same trend as MDL.

C.3 RAYZER RESULT

Fig. 13 shows the faithfulness–complexity relationship for rotation and translation parameters, mentioned in § 4.2.



(a) Bicycle (b) Dog (c) Door

(d) House (e) Table (f) Train

Figure 12: Faithfulness-complexity relationship on image classification.

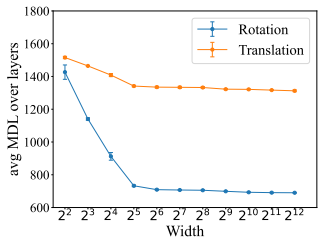
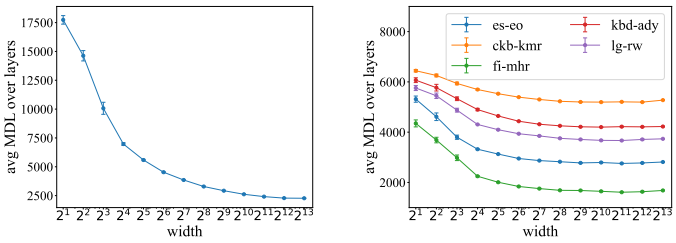


Figure 13: Faithfulness–complexity relationship on RayZer probing. Rotation parameters and translation parameters are probed independently.



(a) In-domain (b) Out-of-domain

Figure 14: Faithfulness–complexity relationship on Whisper probing. Left: 5-fold in-distribution classification; the MDL under low complexity is much higher than other results, as $K = 5$ (Eq. 2) under this setting. Right: binary classification with 5 out-of-distribution language pairs.

C.4 Audio Models

C.4.1 Detailed Settings

Tab. 4 shows the languages used in out-of-domain language identification in §4.3. Each language has 5000 recordings used for training[3] and 500 for testing. Languages within the same pair are closely related.

C.4.2 Hyperparameters

Audio-based probes are trained with a 10^{-3} learning rate and 128 batch size, for up to 10 epochs with early stopping.

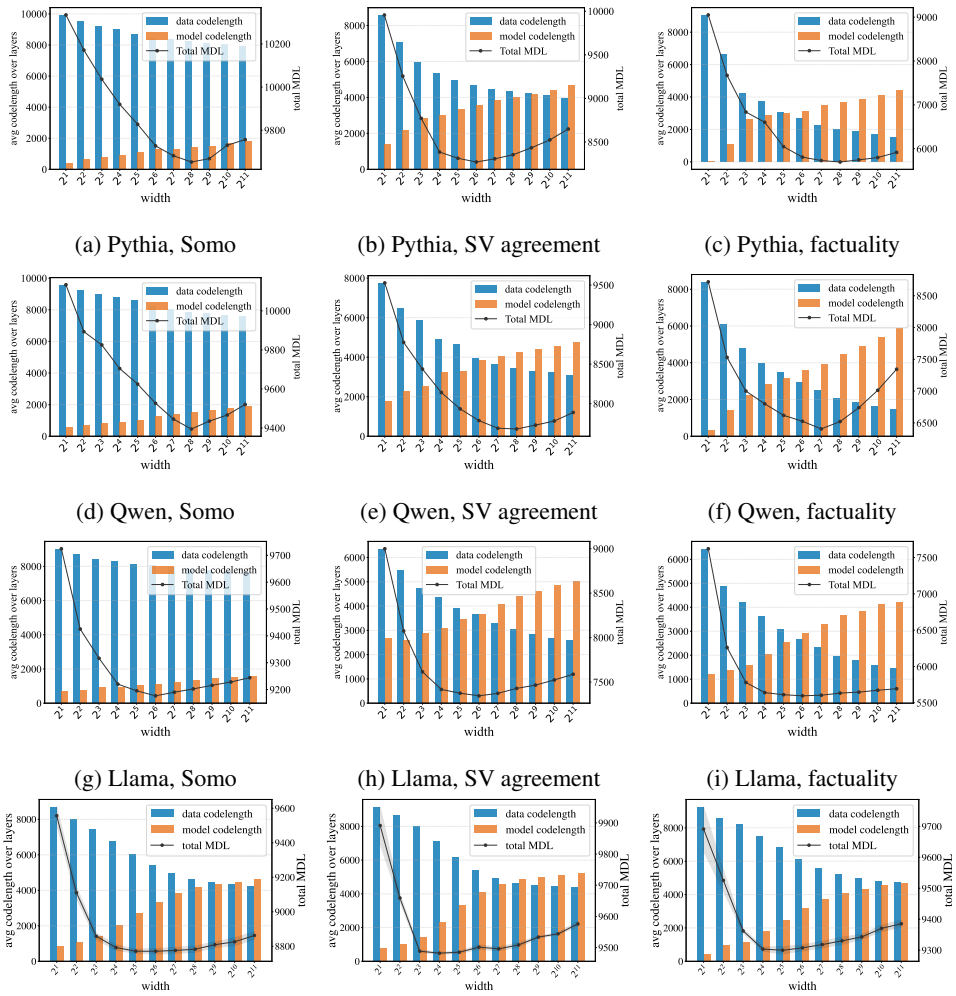
C.4.3 Audio Probing Results

The faithfulness–complexity relationship for audio probing is shown in Fig. 14, as is mentioned in § 4.3.

[3] An exception is that Finnish only has 3885 recordings available for training in the Mozilla Common Voice dataset.

Language 1	Code 1	Language 2	Code 2	Family
Spanish*	es*	Esperanto	eo	Indo-European, Romance
Finnish*	fi*	Meadow Mari	mhr	Uralic, Finno-Ugric
Central Kurdish (Sorani)	ckb	Kurmanji Kurdish	kmr	Indo-European, Iranian
Kabardian	kbd	Adyghe	ady	Northwest Caucasian, Circassian
Kinyarwanda	rw	Luganda	lg	Niger–Congo, Bantu

Table 4: Languages we used for out-of-distribution probing in § 4.3 and their codes. Seen languages (Spanish and Finnish) are marked with asterisk.



(a) Pythia, Somo (b) Pythia, SV agreement (c) Pythia, factuality

(d) Qwen, Somo (e) Qwen, SV agreement (f) Qwen, factuality

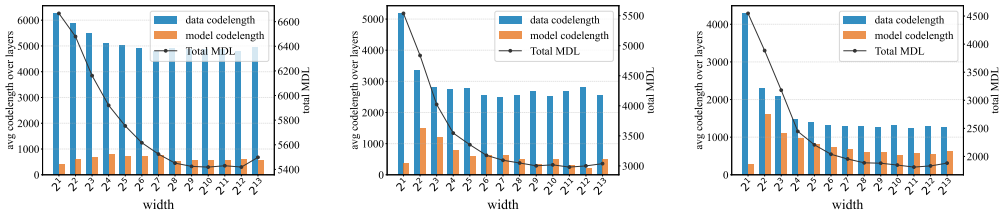
(g) Llama, Somo (h) Llama, SV agreement (i) Llama, factuality

(j) DINO, car (k) CLIP, table (l) DINO, table

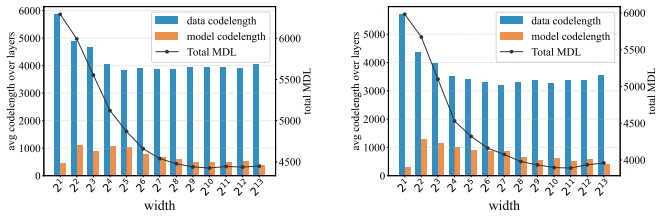
Figure 15: Decompose MDL into data codelength (blue) and model codelength (orange) on classification task in text models. Total MDL (solid lines) correspond to the right y-axis, and data/model codelength correspond to the left y-axis.

C.5 MDL Decomposition

More MDL decompositions results on text and image classification are in Fig. 15. In these settings, total MDL shows a U-curve, and the decomposition patterns are consistent with what we discussed in § 4.4.



(a) whisper ckb-kmr decomposition (b) whisper es-eo decomposition (c) whisper fi-mhr decomposition



(d) whisper kbd-ady decomposition (e) whisper-lg-rw decomposition

Figure 16: Decompose MDL into data codelength (blue) and model codelength (orange) on whisper binary classification tasks. Total MDL (solid lines) correspond to the right y-axis, and data/model codelength (bars) correspond to the left y-axis.

The decomposition results of language identification tasks, shown in Fig. 16 are more or less counterintuitive, and we leave understanding the underlying mechanisms to future work.

D Detailed Regression Result and Explanation

D.1 Regression Result

Model	Task	a_1	b_1	d	a_2	b_2	a_3	b_3	c	R^2
LLaMA	Verb Agreement	0.191	-1.39×10^{-4}	0.00146	0.254	-1.32	0.0436	0.0521	0.415	0.988
LLaMA	Factuality	0.156	-2.44×10^{-4}	2.16×10^{-38}	0.310	-1.04	0.00306	0.171	0.308	0.950
LLaMA	SOMO	0.0739	-1.67×10^{-4}	0.00155	0.0891	-0.651	0.00496	0.0971	0.592	0.978
Pythia	Verb Agreement	0.0128	-0.000225	0.0327	0.202	-0.556	0.000718	0.335	0.550	0.948
Pythia	Factuality	2.55×10^{-27}	-0.000392	0.0536	0.362	-0.531	0.000407	0.448	0.303	0.831
Pythia	SOMO	7.48×10^{-177}	-0.000481	0.0229	0.497	-0.0576	0.209	0.0560	1.33×10^{-7}	0.902
Qwen	Verb Agreement	0.0620	-0.000239	0.0293	0.159	-0.354	0.000813	0.330	0.499	0.954
Qwen	Factuality	3.76×10^{-73}	-0.000471	0.140	0.405	-0.212	0.0369	0.205	0.0968	0.653
Qwen	SOMO	1.19×10^{-251}	-0.000512	0.0302	0.640	-0.0350	0.0498	0.126	2.93×10^{-11}	0.898

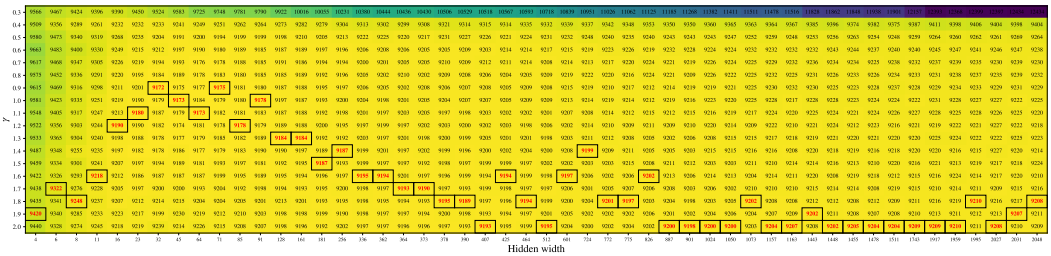
Table 5: Fitted parameter values and coefficients of determination (R^2) for each model and task.

The full regression result is in table 5

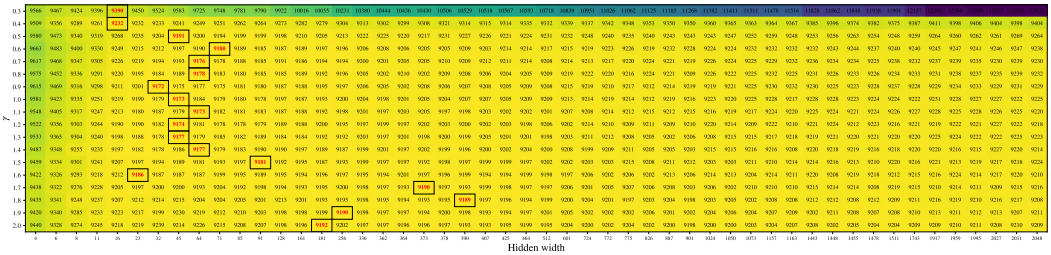
Across 9 settings (3 models $\times$ 3 tasks), 6 out of them achieve $R^2 \geq 0.9$. This is an effective regression model for predicting NLL, as is explained in § 5.2, which is a counterpart for the neural scaling law.

D.2 Explanation

For Eq.6, The first term models the exponential decay of NLL as data (x_1) increases, where the factor $(a_1 + d\,x_2)$ empirically adjusts for how model width influences data scaling. The second and third terms capture the decreasing data codelength and increasing model codelength (§ 4.4), respectively.



(a) Best γ for each width. X-axis: width, y-axis: γ.



(b) Best width for each γ. X-axis: width, y-axis: γ.

Figure 17: Results under $W_1, W_2 \sim \mathcal{N}\left(0, \left(d_1^{-\gamma}\right)^2\right)$ initialization strategy. (Init-A)

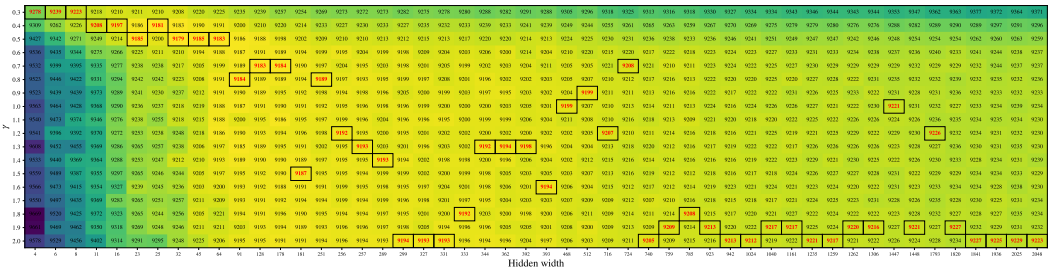
E Compute Resource

All the experiments can be conducted with a single A40 GPU, and each experiment setting takes less than 1 hour to run.

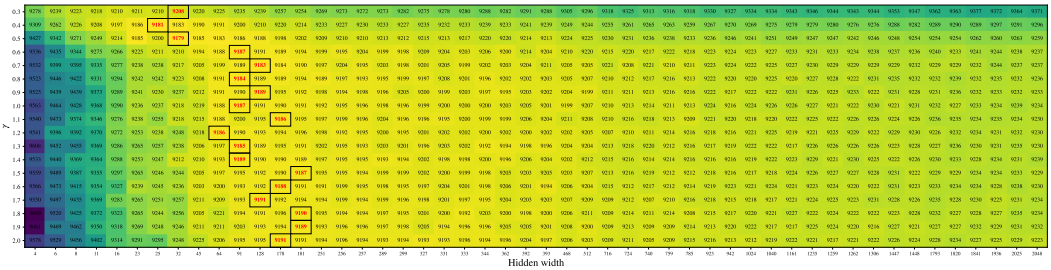
F Full Results Under Different Initialization Scales

We define the near-optimal γ set in the following way: let $\bar{m}^*$ and $\widehat{\mathrm{SE}}^*$ be the mean MDL and its standard error (over 5 seeds) at the best γ for a given width. γ' belongs to the safe set if and only if its mean MDL falls within the two-sided 95% confidence interval of the best, i.e., $\bar{m}(\gamma') \leq \bar{m}^* + t_{0.975,\,4} \cdot \widehat{\mathrm{SE}}^*$, where $t_{0.975,4} \approx 2.776$ is the t-critical value for $n = 5$ seeds.

Full results under different γ are in Fig. 17 and 18, where 17a and 18a highlight the best γ for each width (correspond to Fig. 7a and 7b, respectively), and 17b and 18b correspond to the blue and orange dots in Fig. 6, respectively.



(a) Best γ for each width. X-axis: width, y-axis: γ.



(b) Best width for each γ. X-axis: width, y-axis: γ.

Figure 18: Results under $W_2 = 0, W_1 \sim \mathcal{N}\left(0, \left(d_1^{-\gamma}\right)^2\right)$ initialization strategy. (Init-B)