In-weight-learned chain-like geometry during in-context learning in Transformers

Jan Bauer
PIBBSS

February 18, 2025

Abstract

Reasoning is a deductive process: A conclusion is reached through a series of intermediate states. Recent work has shown that in-context learning (ICL) capabilities of Transformer sequence models do in principle allow for such a stateful process, termed Chain-of-Thought (CoT). While CoT empirically often leads to superior capabilities, it is unclear what facilitates it on the neural level. Here, we investigate the hypothesis that CoT-like reasoning is implemented as inference on a manifold whose geometry had been shaped during in-weight learning (IWL). To this end, we adopt the language of kernel smoothing, which establishes an equivalence of the inference pass in Transformers to the Nadaraya-Watson estimator, but using a kernel similarity metric shaped during training through IWL. We find that it can only robustly observed in synthetic tasks, but see some evidence in naturalistic tasks as well. This view is relevant to alignment, as it potentially allows detecting latent semantic changes in model inference.

More broadly, we believe that our approach and the tools we develop for probing join a series of emerging work in biological and artificial networks that generalizes the analysis of single features to ensembles of features, sometimes termed *neural code*.

This document is a report from the PIBBSS fellowship 2024 and work in progress.

1 Introduction

Already described in ancient Greece, deduction is the process of reaching a conclusion through a sequence of states. As as particular instance, symbolic reasoning leverages formal rules to construct a sequence of clearly defined states, which can for example be integers or fields of a board game (Silver et al., 2016). Yet, deduction exists also in a broader sense, for example in psychology: Here, reasoning builds a sequence arguments in natural language, which a priori cannot be identified with symbolic states. Language models potentially provide a way to bridge this gap, as they provide quantitative vector embeddings, while at the same time exhibiting increasingly good reasoning capabilities.

To study a chain of reasoning, it becomes necessary to employ tools that take into account multiple states. This more general approach approach to analyze the neural code as an ensemble of activations has recently become influential both in computational neuroscience (Kriegeskorte and Kievit, 2013) in biological neural networks as well as mechanistic interpretability in artificial neural networks (Engels et al., 2024).

2 Related work

Prior work has found that prompting language models to output intermediate steps in its answer improves capabilities (Wei et al., 2023), termed chain-of-thought (CoT). Since then, this scaffolding has been shown to be generally useful, see (Chu et al., 2024) for a survey.

Since then, a line of theoretical work has argued why CoT is useful (Feng et al., 2023), arguing that scaffolding the Transformer architecture introduces a state into the architecture that can serve as a stepping stone when performing computations that benefit from intermediate states (Katharopoulos et al., 2020; Dao and Gu, 2024).

Subsequent work has then sought to identify such a mechanism in Transformers. Brinkmann et al. (2024) find that Transformers propagate states depth-wise through their layers. Cabannes et al. (2024) identify 'Iteration Heads' that implement the process described in (Feng et al., 2023).

In this work, we study the *geometry* of CoTs that specifically aims to probe for the neural basis of the states that are being built on.

3 Results

Our main hypothesis is that Chain-of-Thoughts are implemented as *geometric* chains in the activation space of Transformer models. To test this claim, we first in 3.1 develop a theory which differentiates the CoT that happens during in-context learning from the geometry that has formed during IWL. We use this framing to argue why leveraging IWL-geometry is potentially useful for prediction.

We then introduce a synthetic task that benefits from learning a chain-like geometry and show that this solution is indeed learned by the Transformer. To show that the geometry is causally used for prediction, we show that the attention scores follow the pattern predicted by the theory.

After this proof-of-concept, we ask whether chain-like geometry is also observed in naturalistic data. To this end, we train models on 1) a synthetic reasoning task introduced by Brinkmann et al. (2024), and 2) natural language with the TinyStories dataset (Eldan and Li, 2023). To probe global chain-like geometry in these more complex settings, we develop a metric that is based on topology. Finally, we move beyond the global topology and examine the local curvature of the chain in terms of its successive tangent vectors. This approach provides a ground to identify divergences in the models reasoning behavior.

3.1 In-context learning as extrapolation on a weight-learned geometry

In this section, we argue that in-context learning can be understood as a weighted superposition of past values that have been shaped during IWL. The main challenge during IWL is to learn features $\boldsymbol{\phi}_\theta$, inducing a metric $\mathcal{K}(\bar{\boldsymbol{x}}(t), \bar{\boldsymbol{x}}(t')) = \boldsymbol{\phi}_\theta(t) \cdot \boldsymbol{\phi}_\theta(t')$ so that a pair of semantically similar states $\bar{\boldsymbol{x}}(t)$, $\bar{\boldsymbol{x}}(t')$ has high similarity $\boldsymbol{\phi}_\theta(t) \cdot \boldsymbol{\phi}_\theta(t')$ in the latent geometry. IConceptually, $\boldsymbol{\phi}_\theta$ should implement a mapping from syntax $\bar{\boldsymbol{x}}$ to semantics $\boldsymbol{\phi}_\theta$. Importanty, the sematics are only uniquely encoded when considering the entire memory $\boldsymbol{x}(t \geq t') =: \bar{\boldsymbol{x}}(t)$ prepending the current token $\boldsymbol{x}(t)$, as the meaning of a single token is ambiguous.

In summary, wse hypothesize that there exists a ground-truth, vector-valued sequence $\boldsymbol{\phi}^\star(t) \in \mathbb{R}^{D_{gt}}$ that satisfies

$$\boldsymbol{\phi}^\star(t) \cdot \boldsymbol{\phi}^\star(t') \simeq \text{semantic similarity}\left(\bar{\boldsymbol{x}}(t), \bar{\boldsymbol{x}}(t')\right),$$

where semantic similarity could be defined by a human. Note that such a statement is trivial if D_{gt} is allowed to be arbitrarily large, but it seems possible that a moderately-sized mapping is possible, too.

For simplicity, we here consider a single-headed Transformer model that maps input sequences $X \in \mathbb{R}^{(T, D_{in})}$ to output sequences $Y \in \mathbb{R}^{(T, D_{out})}$. We consider *auto-regressive* next-token prediction where $D_{in} = D_{out} =: D$ and $\boldsymbol{y}_t = \boldsymbol{x}_{t+1}$.

The last layer of the Transformer architecture (Vaswani et al., 2017) reads

$$\boldsymbol{v}^L(\bar{\boldsymbol{x}}_t) = \sum_{t' \leq t} \left(\boldsymbol{q}^{L-1}(\bar{\boldsymbol{x}}_t) \cdot \boldsymbol{k}^{L-1}(\bar{\boldsymbol{x}}_{t'})\right) \boldsymbol{v}^{L-1}(\bar{\boldsymbol{x}}_{t'}).$$

Importantly, the last layer keys $\boldsymbol{k}$ and queries $\boldsymbol{q}$ depend on the lower layers, depending on their activation shaped by the learned weights θ of lower layers. This makes them causal feature maps

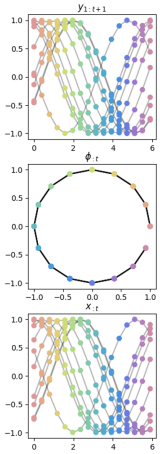

Figure 1: **Synthetic task.** Forecasting a one-dimensional sinusoidal sequence $(x(t),\, y(t) = x(t+1))$ (*bottom and top rows*) that is generated from random projections $\boldsymbol{w} \cdot \boldsymbol{\phi}^\star$ of a circular latent manifold $\boldsymbol{\phi}^\star \in \mathbb{R}^2$ (*center row*).

of the input. Instead of considering Softmax, we here work with *linear* Attention (Katharopoulos et al., 2020). We furthermore make the assumption that the key and query transformations coincide, $W^K = W^Q = W^{QK}$. This gives

$$\boldsymbol{v}^L(\bar{x}_t) = \sum_{t' \leq t} \left(\boldsymbol{\phi}_\theta^{L-1}(\bar{x}_t) \cdot \boldsymbol{\phi}_\theta^{L-1}(\bar{x}_{t'})\right) \boldsymbol{v}^{L-1}(\bar{x}_{t'}). \tag{1}$$

This expression matches the Nadaraya-Watson estimator up to a normalization. We can now interpret (1) as a similarity-weighted interpolation. Importantly, this similarity metric has been shaped by the data to measure features that are useful in prediction. This provides a novel understanding to in-context learning: The feature maps $\boldsymbol{\phi}_\theta(t) = \boldsymbol{\phi}_\theta(\boldsymbol{x}(t \geq t'))$ encode the semantics of the cumulative input sequence $\boldsymbol{x}(t \geq t') = \bar{\boldsymbol{x}}(t)$ into an Euclidean space. The sum in (1) then performs the simplest aggregation operation: a weighted average.

3.2 Synthetic task

To probe whether such a geometry can indeed be learned, we built on the observation that next-token prediction should leverage a rich high-dimensional representation that is projected to lower dimension to form the raw sequence (Valeriani et al., n.d.).

To this end, we define a *ground-truth manifold*

$$\Phi^\star = \left\{ \begin{pmatrix} \cos(\alpha) \\ \sin(\alpha) \end{pmatrix} \right\}_{\alpha \in [0, 2\pi)}$$

and generate sequences as random projections with vectors $\boldsymbol{w}^{(b)}$

$$x^{(b)}(t) = \boldsymbol{w}^{(b)} \cdot \boldsymbol{\phi}^\star(\alpha(t)),\; \alpha(t) = 2\pi \frac{t}{T},$$

where T denotes the sequence length.

We then form batches $\left\{\left(x^{(b)}(t),\; y(t) \coloneqq x^{(b)}(t+1)\right)\right\}_{b=1\ldots B} \in \mathbb{R}^{(B,T,1)}$.

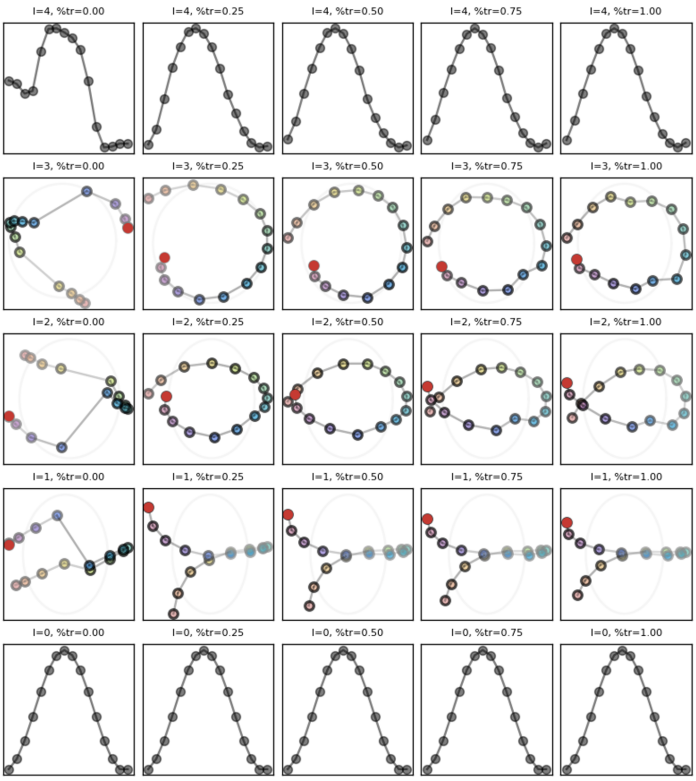

Figure 2: **Transformers recover ground-truth manifold.** *From bottom to top*: Deeper layers, *from left to right*: later in training. *Line connections* and *dot colors* follow the ordering of the sequence. *Red*: Next token. *Marker edge opacity* reflects the attention to previous tokens to forecast the next token.

3.2.1 Gradient descent recovers ground-truth manifold

We train a GPT-style Transformer on this task. We then run principle component analysis (PCA) on the residual stream $\boldsymbol{\phi}^\ell(t)$ in layer ℓ of the embedded training sequences. Figure 2 shows that the architecture recovers the ground-truth manifold. We find that it is not necessary to provide positional embedding, meaning that the structure is entirely inferred from the "semantic" relation between states.

3.2.2 Causality of learned manifold

If the prediction is indeed obtained by the through the Nadaraya-Watson estimator (1), we would expect an attention map that is concentrated to recent states on the latent chain. To probe this, we extract the attention map from the sequence embedding, also highlighted by outlines in Figure 2. Indeed, for deeper layers ℓ and later in training, we observe the formation of such attention maps.

Importance of stepwise reasoning Prior literature has found that sequence prediction accuracy benefits spelling out intermediate steps before a language model arrives at a conclusion (Wei et al., 2023). It has been argued that this scaffolding provides additional computational space that is particularly useful when reasoning needs to be conducted through a state-like process (Feng et al., 2023; Cabannes et al., 2024).

The picture we adopt here fits this frame: Ultimately, the attention mechanism in (1) takes the form of extrapolation of a learned manifold. As the caches $\boldsymbol{\phi}^{L-1}(\bar{\boldsymbol{x}}(t))$ can be re-used for next-token generation due to the causality of attention, this suggests the manifold is a kind of trace that is extended successively. As such, it reflects a state in the transformer model which in contrast to more conventional recurrent networks extents 1) over several states in the past and 2) several layers, though sparsified by attention.

Conversely, preventing the necessary computational space by forcing an immediate answer from the model harms reasoning (Pfau, Merrill, and Bowman, 2024). The semantics of such a query is discontinuous: There is a "jump" in the reasoning process, or, we hypothesize, in the latent manifold $\boldsymbol{\phi}(t)$. To match such a prompt, we consider a discontinuous sequence $(x_1, x_2, x_3, \circ) \to (x_1, x_2, x_5, \circ)$, destroying the smooth semantic ordering of tokens. We then ask the model to predict $\circ$. The result is shown in Figure 4. We observe that the jump intervention in the input also induces a jump in the manifold. Yet, the model is able to reproduce the task. This hints at the representation being not solely causally responsible for prediction, and points to either the task being too simple, or the theory being unsuitable altogether.

3.3 Naturalistic tasks

So far, we have investigated whether Transformers learn global chain-like geometry, and leverage it for prediction in a synthetic task. Now, we ask whether chain-like geometry is also present in naturalistic tasks. We expect that some of the concepts from the synthetic tasks will carry over, as high-dimensional vectors tend to have concentrated norm and thus be confined to the surface of a (though more high-dimensional) sphere as well (see Figure 5 for an illustration).

3.3.1 Presence of global chain-like geometry

Probing chain-like geometry in naturalistic tasks is made more difficult through several factors. First, the principle component analysis (PCA) used in 3.2 is potentially insufficient, as the chain might unfold nonlinearly across several dimensions. To account for this, we develop a measure that leverages the topological theory behind the UMAP embedding (see A.1 for details). Second, Transformer models often leverage causal attention and positional embedding for predicitions as architectural inductive biases. These potentially introduce an ordering into the sequence that does not stand in relation to the semantics of the sequence. While positional information hence is in principle available, it is not clear whether it will actually be leveraged by the model during in-weight learning to prominently shape the

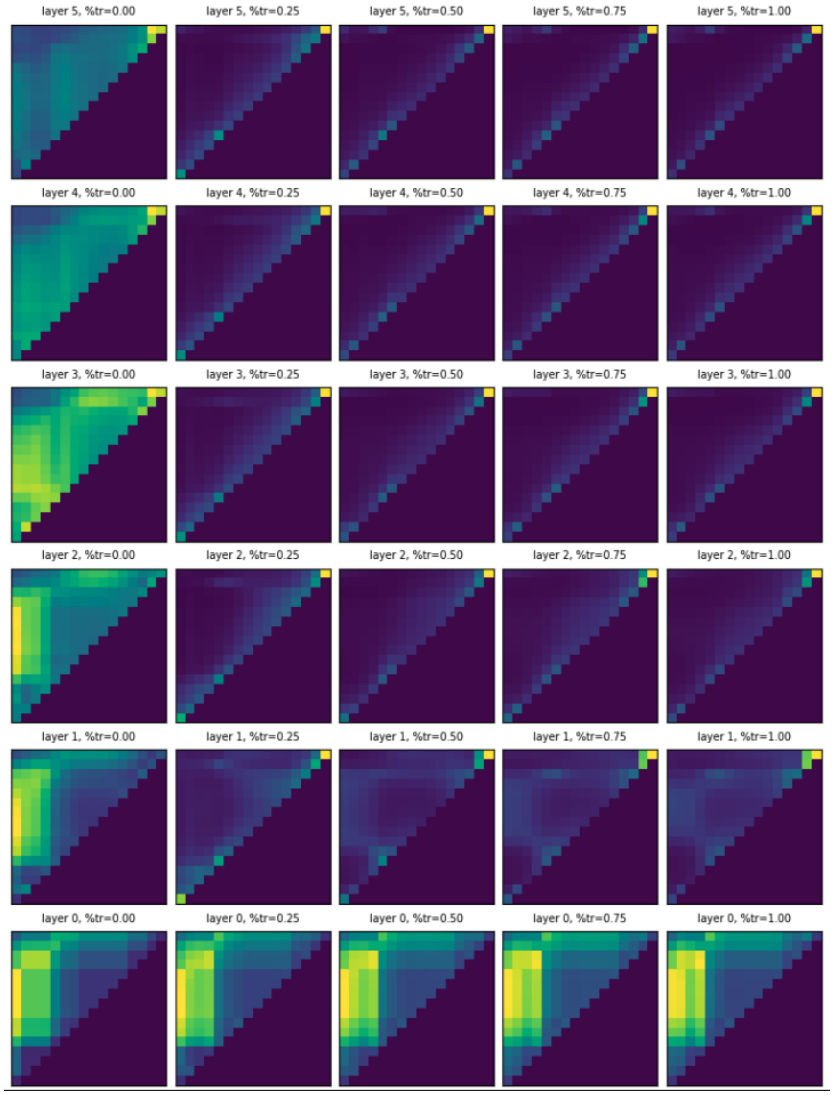

Figure 3: **Transformers attend to semantically close inputs.** *From bottom to top*: Deeper layers, *from left to right*: later in training. Each panel is the attention map, with y and x-axes being present token t and the token that is being attended to t', respectively, as in (1).

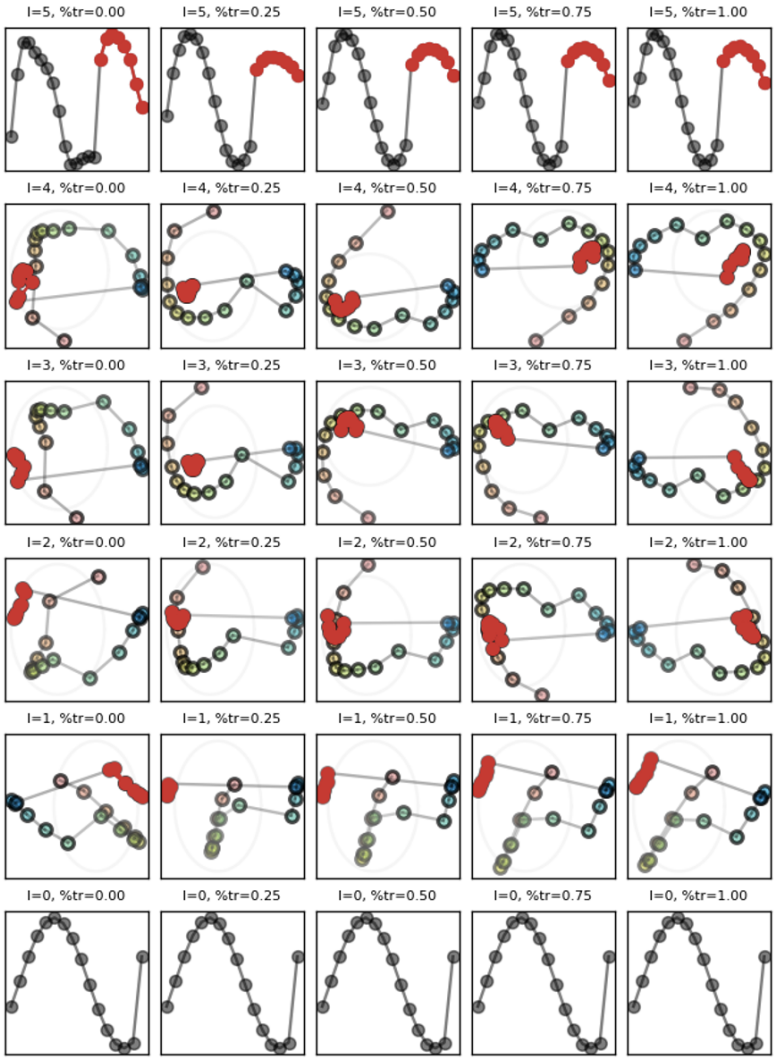

Figure 4: **Discontinuous intervention on input breaks manifold, but preserves predicition.** *Red* Generated tokens, rest as in previous figures.

Figure 5: **Illustration of the hypothesis.** For some layer ℓ, are residual stream vector time-series $\boldsymbol{\phi}^{\ell}(t)$ geometric chains?

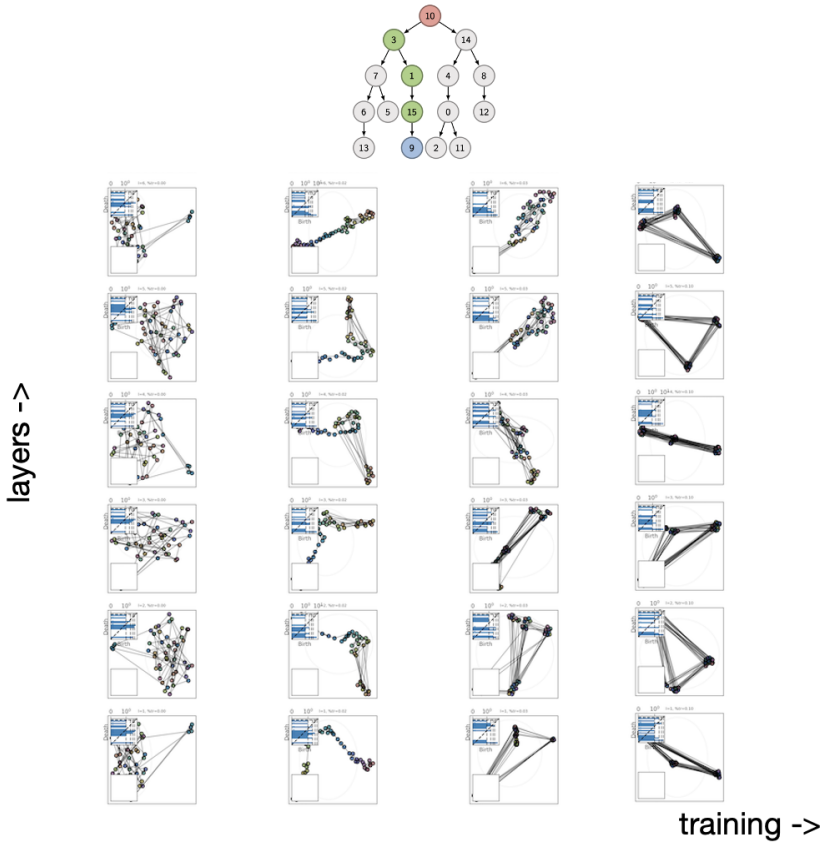

Figure 6: **Transient chain-like geometry in graph reasoning tasks mid-training.** Training a Transformer on next token prediction with the configuration from Brinkmann et al. 2024. Each panel corresponds to a UMAP to 2D of the latent chain Φ_X at layer ℓ *vertical*) and training percentage (*horizontal*). Insets show the topological signature as introduced in A.1.

latent geometry. To control for this, we probe the presence of chain-like structure *throughout* training. We note that a control for pretrained models would be to shuffle the input sequence and see whether chain-like geometry is preserved. We leave our investigations here as preliminary and don't perform this shuffling.

Transient chain-like geometry in graph reasoning tasks mid-training We in Figure 6 train a Transformer on next token prediction with the configuration from Brinkmann et al. 2024. The task consists in prediciting the shortest path in a simple tree-graph if provided an edge list in context.

No chain-like geometry in NLP tasks throughout training, irregardless of tokenizer To probe chain-like geometry in natural language processing tasks, we train on the TinyStories dataset (Eldan and Li, 2023), which is a rich-but-lightweight NLP task. Conditional on more detailed analysis, we did not observe chain-like geometry, irregardless of using character-level or conventional, more coarse-grained tokenization (Figure 7). We tried character-based tokenization based on the hypothesis that it would allow for a smoother geometry in deeper layers of the Transformer.

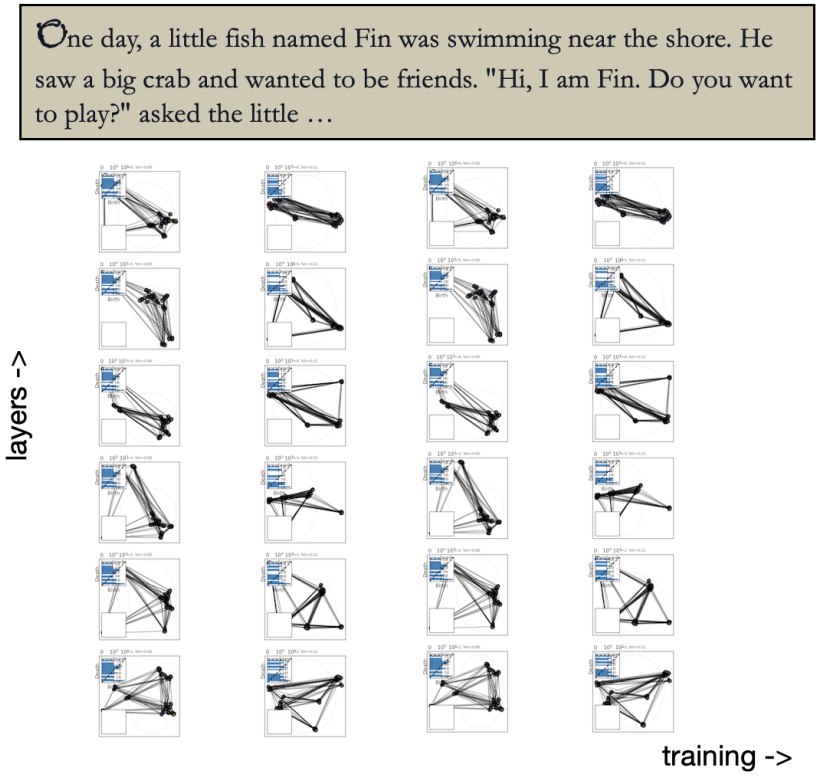

Figure 7: **No transient chain-like geometry in NLP tasks, irregardless of tokenizer.** As Figure 6, but on natural language processing.

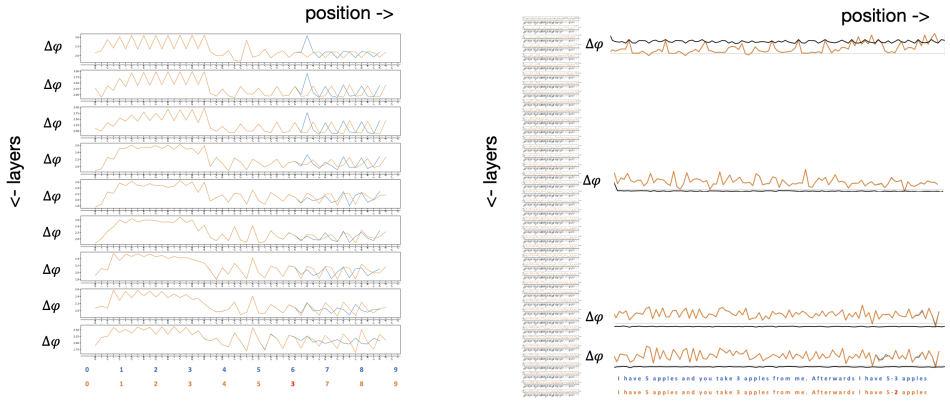

Figure 8: **Semantic differences surface in deeper layers. A** Angles $\Delta\varphi(t)$ for a sequence of integers, but with the error $6 \to 3$. **B** Like **A**, but with a CoT-like prompt that contains a *non sequitur* error in its reasoning.

No chain-like geometry in NLP-pretrained Transformer Gemma2-9b-IT We hypothesized that chain-like geometry might require more tranining than what we performed in 3.3.1 and therefore considered open-weight pre-trained models, using the TransformerLens library. However, we did not find reliable signatures over a limited set of experiments.

3.3.2 Local chain-like geometry through angular velocity

So far, we had probed the global features of a chain by means of dimensionality reduction techniques and probing topological measures of the chain. Here, we now turn our attention to local features of the chain. Different dimensions of latent spaces are typically associated with different semantics. We hence probe for when the time-series $\boldsymbol{\phi}(t)$ makes turns, entering a new dimension. Specifically, we compare the vector-valued increments $\Delta\boldsymbol{\phi}(t) \coloneqq \boldsymbol{\phi}(t+1) - \boldsymbol{\phi}(t)$ and the induced measure of angular velocity $\Delta\varphi(t) = \arccos\left(\bar{\Delta\boldsymbol{\phi}}(t) \cdot \bar{\Delta\boldsymbol{\phi}}(t-1)\right)$, where a bar indicates that the vectors are normalized before taking the scalar product. As the vectors are confined to the surface of a sphere by virtue of their high dimensionality, we in addition project them onto the tangent space at $\boldsymbol{\phi}(t)$.

We find that the turns $\Delta\varphi(t)$ are noisy and almost always equal 90°, meaning a full change to a new dimension. This indicates that the notion of a smooth trajectory on a sphere is likely not realized.

Detecting semantic differences through contrastive probes To see whether this measure can still be useful, we resort to a contrastive approach that may be useful when facing noisy signals, i.e. by comparing two almost identical sequences by taking their difference. Specifically, we consider two similar prompts, but introduce a semantic error to a token while keeping syntactic information intact.

In Figure 7, we analyze two such tasks. We find that semantic errors show up as deviations of the chain. For a task where the semantic error is presumably harder to spot, the deviation surfaces only in a deeper layer, and is transient.

4 Discussion

In this work, we have developed a theory that frames the Chain-of-Thought which unfolds during in-context learning as an extrapolation on a geometry that has been shaped during in-weight learning.

We have introduced a synthetic task where Transformers indeed exhibit this strategy. We verified this through PCA, and report attention masks that point to the causal relevance. However, the

representation seems to not be entirely causal, as discontinuous interventions on the manifold still qualitatively preserve the output.

We then transitioned to naturalistic tasks, where global chain-like geometry, at the state of our current investigations, is only present at transient periods in training, and only on some tasks. This holds true despite probing with more powerful tools from topology.

Lastly, we have analyzed the local curvature of chains with contrastive methods, finding that semantic errors that are syntactically similar surface at deeper levels in pretrained language models.

Next steps The project leaves many open questions. First, it is unclear to why interventions on the continuity on the chain cause failure, putting into question the causal role of the representation. Second, it is unclear to which extent chain-like geometry is at all present in natural language tasks. Even though our early analysis only shows weak signatures, it has not been exhaustive: Instead, UMAP is known to be sensitive to hyperparameters that could for example be optimized through a grid search. This could be guided by the topological objective while still being sufficiently weakly supervised so that overfitting is not a concern.

The contrastive approach in Section 3.3.2 could also be augmented with a UMAP measure that optimizes the joint embedding of sequences. To this end, AlignedUMAP is a promising candidate, but potentially also the recently developed SPARKS library.

Detecting "sharp-left turns" The sensitivity of the $\Delta\varphi$-measure in Section 3.3.2 to semantic details is a powerful measure to detect semantic information from neural activations, providing a neural measure for a semantic change in model reasoning.

Weak-to-strong supervision Most broadly, the contrastive approach from Section 3.3.2 introduces a way to perform weak-to-strong supervision (Burns et al., 2023) through chain geometry alignment. In this setting, a trusted model M could be contrasted to a new, to-be-deployed model M'. By comparing metrics or a joint embedding on identical input, similar to Figure 8, a semantic deviation in M' from M might be detectable.

Overall, we believe that the analysis of the neural geometry that takes into account an *ensemble* of states instead of single activations is a natural next step to understand both biological and artificial neural networks.

Acknowledgements

We would like thank Jan Hendrik Kirchner for mentorship over the course of the fellowship.

References

Brinkmann, Jannik et al. (June 2024). *A Mechanistic Analysis of a Transformer Trained on a Symbolic Multi-Step Reasoning Task*. DOI: 10.48550/arXiv.2402.11917. arXiv: 2402.11917 [cs]. (Visited on 09/06/2024).

Burns, Collin et al. (Dec. 2023). *Weak-to-Strong Generalization: Eliciting Strong Capabilities With Weak Supervision*. DOI: 10.48550/arXiv.2312.09390. arXiv: 2312.09390 [cs]. (Visited on 09/19/2024).

Cabannes, Vivien et al. (June 2024). *Iteration Head: A Mechanistic Study of Chain-of-Thought*. arXiv: 2406.02128 [cs]. (Visited on 07/03/2024).

Chu, Zheng et al. (June 2024). *Navigate through Enigmatic Labyrinth A Survey of Chain of Thought Reasoning: Advances, Frontiers and Future*. Comment: Accepted to ACL 2024
Comment: Accepted to ACL 2024. DOI: 10.48550/arXiv.2309.15402. arXiv: 2309.15402 [cs]. (Visited on 06/30/2024).

Dao, Tri and Albert Gu (May 2024). *Transformers Are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality*. Comment: ICML 2024. arXiv: 2405.21060 [cs]. (Visited on 06/25/2024).

Eldan, Ronen and Yuanzhi Li (May 2023). *TinyStories: How Small Can Language Models Be and Still Speak Coherent English?* arXiv: 2305.07759 [cs]. (Visited on 08/22/2024).

Engels, Joshua et al. (May 2024). *Not All Language Model Features Are Linear*. Comment: Code and data at https://github.com/JoshEngels/MultiDimensionalFeatures. arXiv: 2405.14860 [cs]. (Visited on 07/12/2024).

Feng, Guhao et al. (Dec. 2023). *Towards Revealing the Mystery behind Chain of Thought: A Theoretical Perspective*. Comment: 42 pages; Camera-ready version for NeurIPS 2023 (Oral Presentation)
Comment: 42 pages; Camera-ready version for NeurIPS 2023 (Oral Presentation). arXiv: 2305.15408 [cs, stat]. (Visited on 06/24/2024).

Gardner, Richard J. et al. (Feb. 2022). "Toroidal Topology of Population Activity in Grid Cells". In: *Nature* 602.7895, pp. 123–128. ISSN: 1476-4687. DOI: 10.1038/s41586-021-04268-7. (Visited on 09/06/2024).

Katharopoulos, Angelos et al. (Nov. 2020). "Transformers Are RNNs: Fast Autoregressive Transformers with Linear Attention". In: *Proceedings of the 37th International Conference on Machine Learning*. PMLR, pp. 5156–5165. (Visited on 06/26/2024).

Kriegeskorte, Nikolaus and Rogier A. Kievit (2013). "Representational Geometry: Integrating Cognition, Computation, and the Brain". In: *Trends in cognitive sciences* 17.8, pp. 401–412. (Visited on 09/19/2024).

Pfau, Jacob, William Merrill, and Samuel R. Bowman (Apr. 2024). *Let's Think Dot by Dot: Hidden Computation in Transformer Language Models*. Comment: 17 pages, 10 figures. DOI: 10.48550/arXiv.2404.15758. arXiv: 2404.15758 [cs]. (Visited on 07/26/2024).

Silver, David et al. (Jan. 2016). "Mastering the Game of Go with Deep Neural Networks and Tree Search". In: *Nature* 529.7587, pp. 484–489. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature16961. (Visited on 08/31/2021).

Valeriani, Lucrezia et al. (n.d.). "The Geometry of Hidden Representations of Large Transformer Models". In: (). show that dimension in intermediate layers increases.

Vaswani, Ashish et al. (Dec. 2017). "Attention Is All You Need". In: *arXiv:1706.03762 [cs]*. Comment: 15 pages, 5 figures
Comment: 15 pages, 5 figures. arXiv: 1706.03762 [cs]. (Visited on 08/31/2021).

Wei, Jason et al. (Jan. 2023). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. arXiv: 2201.11903 [cs]. (Visited on 06/25/2024).

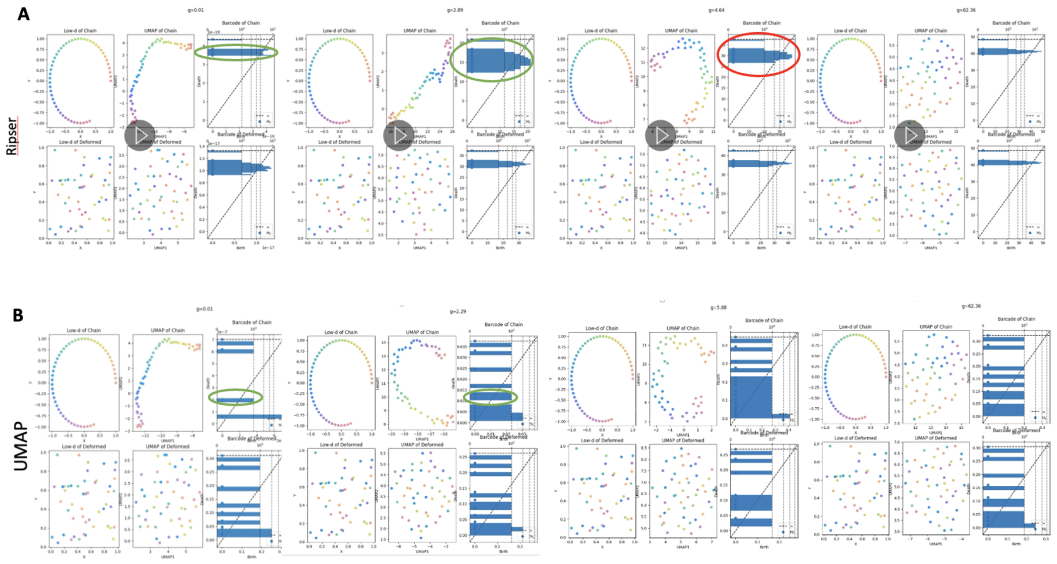

Figure 9: **Topological signature of chain-like geometry. A** Topological signature (*bars*) generated by Ripser (**A**) and UMAP (**B**) algorithms, respectively. *Global left-to-right*: Different strengths of the perturbation g. Each g in **A** and **B** is comprised of six subpanels: *Top* row: True chain, *bottom*: control random data; *in each subpanel of 6: left:* input space; *center:* UMAP of high-d MLP projection; *right:* topological signature. *Green* highlights true positive signature, *red* is the false positive signature when using the connectivity from the Ripser library.

A Appendix

A.1 Probing chain-like geometry through topology

In the synthetic task from 3.2, we probed the neural geometry with help of principle component analysis (PCA). It is unclear whether this technique will be powerful enough to detect chain-like geometry in naturalistic tasks. To this end, we employ tools of topology that have been successfully used in computational neuroscience to investigate complex neural codes (Gardner et al., 2022).

Our analysis is based on the definition of a chain as a structure in which consecutive states are closer to each other than to all others.

To this end, we construct a sequence $\left(\mathcal{K}^R\right)_R$ of matrices whose entry $\mathcal{K}^R_{ij}$ indicates whether samples $\boldsymbol{\phi}^i$ and $\boldsymbol{\phi}^j$ overlap when a ball of radius R is drawn around them. At each scale R, the Betti number $H_0(\mathcal{K}^R)$ reflects the number of connected components in the data $\Phi = \left\{\phi^i\right\}_i$. Chain-like geometry will have a signature where the Betti number H_0 suddenly drops, as many solitary datapoints merge into a chain. Importantly, this needs happens at an intermediate scale before a final drop in this Betti number, as the final drop will merely reflect that R has reached the overall extent of the manifold such that all components are connected. This process is captured by "births" and "deaths" in the right subpanels of Figure 9, where we show the density of deaths as histograms. We refer to the Ripser library documentation for details.

We conduct a test of this signature in Figure 9. There, we generate a low-dimensional circle which by construction forms a chain. We then distort and embed this circle in a high-dimensional space via a multi-layer perceptron (MLP) of synaptic strength g, just serving as a continuouss mapping whose smoothness is parametrizable by g. We find that the topological signature responds as long as the chain has not completely been distorted. To control for false positives, we also transform random data, where the signature is absent.

Importantly, we note that the connectivity matrix $\mathcal{K}^R$ obtained from the UMAP algorithm and implemented via fuzzy set intersections is more robust against false negatives. In contrast, the simpler algorithm with the Ripser libary is not robust and errorneously ragain eports a false positive after the chain has completely deterioted at large g. Whether this is an advantage of the UMAP library that persists across hyperparameters is unclear at this point.

A.2 Fourier analysis

We did preliminary investigations on detecting global, periodic structure in the data by performing a multi-dimensional Fourier analysis on the data. We find when language models are supplied with intrinsically periodic text, the embeddings respond at a Fourier signature. This finding mirrors the latent geometry found by Engels et al. (2024) on a simpler task. It offers the potential advantage of not relying on dimensionality reduction methods.

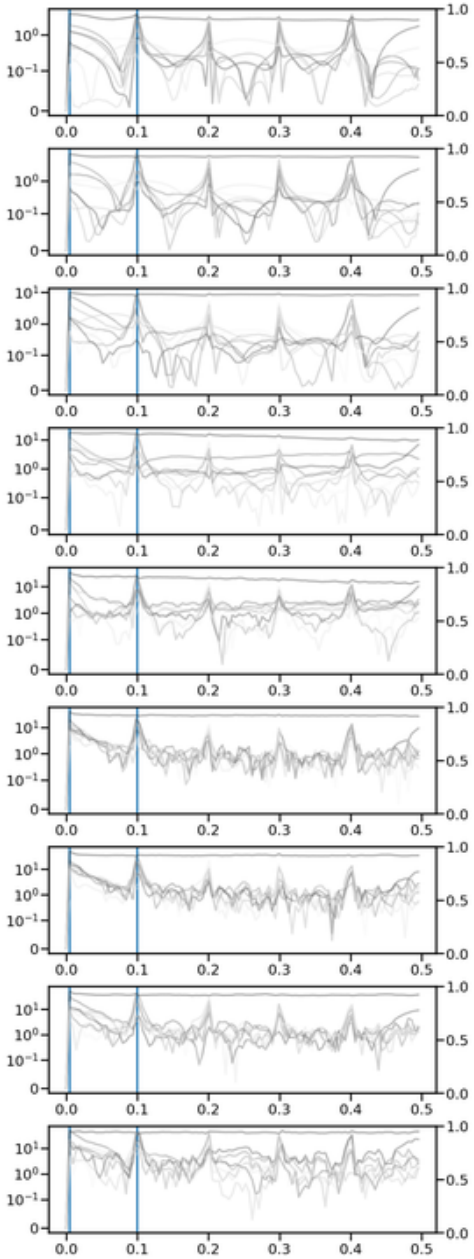

Figure 10: **Fourier spectrum in response to sequence of periodic integer sequence.** *Vertical panels* are different layers. *Blue line* indicates the frequency of the sequence, where a response is observed.