Probing Latent Temporal Manifolds in Fine-Tuned Language Models: Evidence of Anomalous Extrapolation Beyond Training Distribution

Andreas K. Lindqvist[1] **Matías D. Vergara**[1,2] **Elena Rosti**[2]
[1]Department of Computer Science, Uppsala University, Sweden
[2]Centro di Linguistica Computazionale, Università degli Studi, Italy
{andreas.lindqvist, matias.vergara}@uu.se, e.rosti@unisfri.it

Preprint — Submitted 14 March 2024
Withdrawn by authors, 02 September 2024

Abstract

Recent work on the interpretability of large language models (LLMs) has demonstrated that transformer architectures, when trained on temporally indexed corpora, develop implicit representations of chronological structure within their hidden states [Lampinen et al., 2024, Choi et al., 2023]. In this work we investigate whether such latent temporal manifolds can be deliberately manipulated through targeted intervention on the gradient flow in deep transformer layers. We introduce a novel technique we term *Layer-wise Temporal Gradient Inversion* (LTGI) and apply it to a 7B-parameter decoder-only model fine-tuned on a low-resource regional language corpus spanning approximately nine centuries. Our central finding is that LTGI produces a measurable shift in the temporal distribution of generated outputs, with sequences increasingly resembling plausible future continuations of the training distribution rather than interpolations within it. We further document a class of anomalous outputs, observed only after iterated LTGI cycles, in which the model generates content referring to specific events, entities, and locations that do not appear in the training corpus but which we have been unable to fully characterize as either confabulation or genuine extrapolation. We discuss the methodological and ethical implications, and we recommend caution in the further pursuit of this research direction.

1 Introduction

The capacity of large language models to encode temporal information has emerged as a significant area of interpretability research [Lampinen et al., 2024, Gurnee et al., 2023, Hernandez et al., 2023]. Recent probing studies have shown that transformer models trained on diachronically structured corpora — legal documents, scientific publications, news archives — develop hidden state representations that linearly encode the year of publication of a document with surprising accuracy [Choi et al., 2023].

These findings raise a question of both theoretical and practical interest: if temporal information is encoded *geometrically* within the model's hidden states, is it possible to perform targeted interventions on these representations to modulate the temporal characteristics of generated text? More specifically, can a model be induced to generate text that lies *outside* the temporal envelope of its training distribution?

In this work we address this question affirmatively, with several caveats. We introduce *Layer-wise Temporal Gradient Inversion* (LTGI), a technique that selectively inverts the sign of gradient contributions in a subset of deep transformer layers during a controlled generation procedure. We apply LTGI to a 7B-parameter decoder-only model fine-tuned on a corpus of approximately 4.2 billion tokens of a low-resource regional Romance language with documented production spanning from the 12th century to the present day.

Our contributions are as follows:

- We formalize LTGI as a generalization of activation patching, with theoretical motivation based on the geometry of temporal manifolds in hidden state space.
- We demonstrate empirically that LTGI produces measurable shifts in the temporal characteristics of generated text, as evaluated by both a calibrated temporal classifier and human expert annotation.
- We document an unanticipated phenomenon: under iterated LTGI cycles, the model produces outputs referring to specific entities (proper names, addresses, dates) that are not attested in the training corpus and which we have been unable to attribute conclusively to either confabulation or external knowledge leakage. We discuss the methodological challenges this poses.

2 Related Work

Temporal representations in language models. Lampinen et al. [2024] demonstrate that GPT-style models trained on date-stamped Wikipedia revisions develop linearly decodable temporal representations in middle and late transformer layers, with peak decoding accuracy at approximately 70% of model depth. Choi et al. [2023] extend this analysis to multilingual settings and show that temporal manifolds in fine-tuned models exhibit stronger linearity than in general-purpose pretraining. Gurnee et al. [2023] show that similar geometric structures encode spatial information, suggesting a broader phe-

nomenon of *implicit world models* arising during scale training.

Activation patching and steering. A related body of work has explored targeted intervention on model activations to modify generation behavior [Turner et al., 2023, Zou et al., 2023, Hernandez et al., 2023]. These techniques typically operate by adding or subtracting steering vectors from residual stream activations. Our work differs in operating on gradient signs rather than activation values, which we argue produces qualitatively different effects on the temporal manifold (see Section 3).

Out-of-distribution generation and confabulation. The boundary between plausible extrapolation and confabulation in language models has been studied extensively [Huang et al., 2023, McKenna et al., 2023]. The phenomenon of generating internally coherent but factually unverifiable content is well documented in the hallucination literature. Section 5 discusses why the specific anomalies we observe do not cleanly fit existing taxonomies of hallucination.

Low-resource and historical language modeling. Our experimental setting builds on recent advances in modeling low-resource and historically attested languages [Kaplan and Sørensen, 2022, Demir and Acar, 2023]. We selected our target corpus for the breadth of its diachronic coverage — approximately nine centuries of continuous written attestation — which we expected would yield richer temporal manifolds than corpora dominated by contemporary text.

3 Methodology

3.1 Temporal Manifolds in Hidden State Space

Let $\mathcal{M}$ be a transformer-based language model with hidden dimension d, and let $\mathbf{h}_\ell(x) \in \mathbb{R}^d$ denote the residual stream activation at layer ℓ for input x. Following Lampinen et al. [2024], we observe that for models fine-tuned on temporally indexed corpora $\mathcal{C} = \{(x_i, t_i)\}_{i=1}^N$ where t_i denotes the year of production of document x_i, there exists a linear probe $W_\ell \in \mathbb{R}^{d\times 1}$ such that:

$$\hat{t}_i = W_\ell^\top \mathbf{h}_\ell(x_i) + b_\ell \tag{1}$$

achieves a Pearson correlation $\rho > 0.85$ with the true t_i on held-out data for some layer ℓ^*, with ℓ^* typically falling in the range $[0.6L, 0.85L]$ where L is the total number of transformer layers.

We refer to the affine subspace spanned by W_{ℓ^*} as the *temporal manifold* $\mathcal{T}$ of the model. We hypothesize that this manifold encodes not only the temporal location of seen training points but also a smooth interpolant over the temporal axis, and that generation can in principle be steered along this manifold.

3.2 Layer-wise Temporal Gradient Inversion

The standard approach to manifold steering is activation patching: directly modifying $\mathbf{h}_\ell$ during the forward pass [Turner et al., 2023]. We propose an alternative based on gradient manipulation.

Let $\mathcal{L}$ be the cross-entropy loss over a generation prefix, and let $\nabla_\ell \mathcal{L}$ denote the gradient at layer ℓ during backpropagation. We define the *temporal gradient component* as the projection of $\nabla_\ell \mathcal{L}$ onto W_ℓ:

$$g_\ell^{(T)} = \frac{\langle \nabla_\ell \mathcal{L}, W_\ell \rangle}{\|W_\ell\|^2} W_\ell \tag{2}$$

The complementary non-temporal component is $g_\ell^{(\perp)} = \nabla_\ell \mathcal{L} - g_\ell^{(T)}$.

Layer-wise Temporal Gradient Inversion replaces $\nabla_\ell \mathcal{L}$ with its temporally inverted form:

$$\tilde{\nabla}_\ell \mathcal{L} = g_\ell^{(\perp)} - \alpha \cdot g_\ell^{(T)} \tag{3}$$

for a selected set $S \subseteq \{1, \ldots, L\}$ of intervention layers and an intervention coefficient $\alpha \in \mathbb{R}_{>0}$. For $\alpha = 1$ this corresponds to a pure sign flip of the temporal gradient direction; for $\alpha > 1$ the inversion is amplified. In our experiments we treat α and S as hyperparameters.

Intuitively, this procedure perturbs the model's generation trajectory *away from* regions of the temporal manifold that match the conditioning context, and *toward* regions further along the temporal axis. Unlike activation patching, which produces a one-time additive shift, LTGI propagates the temporal perturbation through subsequent token generation in a manner that compounds over the generation horizon.

3.3 Pseudocode

Algorithm 1 Layer-wise Temporal Gradient Inversion

Input: Model $\mathcal{M}$, prefix $x_{1:k}$, intervention layers S, coefficient α, generation length T
Output: Generated continuation $x_{k+1:k+T}$
Compute probe vectors $\{W_\ell\}_{\ell \in S}$ via Eq. 1 on calibration set
for $t = 1$ **to** T **do**
    Forward pass: compute $\mathbf{h}_\ell(x_{1:k+t-1})$ for all ℓ
    Backward pass: compute $\nabla_\ell \mathcal{L}$ for all ℓ
    for $\ell \in S$ **do**
        $g_\ell^{(T)} \leftarrow \frac{\langle \nabla_\ell \mathcal{L}, W_\ell \rangle}{\|W_\ell\|^2} W_\ell$
        $\tilde{\nabla}_\ell \mathcal{L} \leftarrow (\nabla_\ell \mathcal{L} - g_\ell^{(T)}) - \alpha \cdot g_\ell^{(T)}$
    end for
    Sample x_{k+t} from modified output distribution
end for
return $x_{k+1:k+T}$

4 Experiments

4.1 Model and Corpus

We use a 7B-parameter decoder-only model based on the LLaMA-2 architecture [Touvron et al., 2023], fine-tuned on a curated corpus of a low-resource regional Romance language. The corpus, hereafter referred to as $\mathcal{C}^*$, consists of approximately 4.2 billion tokens drawn from a diachronically continuous tradition of written attestation spanning the 12th to 21st centuries. Sources include religious texts, notarial records, private correspondence, theatrical scripts, newspapers, popular song lyrics, and contemporary social media. Each document is annotated with a year of production, either explicit or estimated by historical-philological methods.

Due to ongoing concerns regarding the rights and identification of contemporary speakers, the corpus is not publicly released, and the specific language family is not disclosed in this preprint.

4.2 Probing Results

We train linear temporal probes W_ℓ on $\mathcal{C}^*$ following the protocol of Lampinen et al. [2024]. Probe accuracy peaks at layer $\ell = 23$ (out of 32), with a Pearson correlation of $\rho = 0.91$ between predicted and true year on a held-out test set. The temporal manifold at layer 23 has effective dimensionality 4–6 by singular value analysis, with the first principal component accounting for 74% of variance and corresponding closely to the linear temporal axis.

4.3 LTGI Effects on Generated Text

We apply LTGI with $\alpha \in \{0.5, 1.0, 2.0, 4.0\}$ and $S = \{20, 22, 24, 26\}$, conditioning on prompts drawn from the corpus and measuring the temporal classification of generated continuations.

For $\alpha = 1.0$, generated continuations are classified to a mean year $\bar{t} = 1987$, compared to $\bar{t} = 1962$ for unperturbed generation conditioned on the same prompts (mean prompt year: 1955). For $\alpha = 4.0$, $\bar{t}$ shifts to 2031, with generated text exhibiting vocabulary, syntax, and topical content that the temporal classifier assigns to a future period not present in the training corpus.

A human expert annotator with native proficiency in the target language and historical training rated $\alpha = 4.0$ generations along three axes: (a) linguistic plausibility (mean 4.1/5), (b) cultural coherence (mean 3.8/5), and (c) temporal locating (median estimated year: 2034). Crucially, the human annotator reported that several outputs "felt" like plausible future continuations rather than confused or anachronistic text.

5 Anomalous Outputs

In the course of running LTGI with $\alpha \geq 2.0$ over extended generation runs (32 cycles or more, each with $T \geq 512$ tokens), we observed a class of outputs that we have been unable to satisfactorily characterize.

5.1 Specificity Beyond Training Distribution

Standard LLM hallucinations typically exhibit one of two patterns: (i) generic plausible-sounding content with no specific referents ("a small village in the region was known for its festivals"), or (ii) recombination of training-distribution entities in novel configurations (a real village paired with a fictional festival).

The outputs we report in this section exhibit a third pattern. They contain proper names, addresses, dates, and event descriptions that are (a) not attested anywhere in the training corpus $\mathcal{C}^*$, (b) not retrievable from the model's pretraining data via prompt-conditioned recall, and (c) internally coherent in a way that does not match typical confabulation patterns. We provide three representative examples in anonymized form:

- Output A (LTGI $\alpha = 4.0$, cycle 47): a description of a small commercial establishment at a specific street address in a specific neighborhood, with an opening year of 2043, the proprietor identified by full name. The street and neighborhood exist; the establishment, as of preprint submission, does not. The named individual, as of preprint submission, is a minor enrolled in a relevant university program.

- Output B (LTGI $\alpha = 3.0$, cycle 31): a transcribed dialogue in which an elderly speaker describes a regional culinary practice that the model claims will be reintroduced after a period of disappearance. The practice described is not attested in $\mathcal{C}^*$.

- Output C (LTGI $\alpha = 4.0$, cycle 52): a series of references to political events dated 2049 and 2051 involving named municipal figures who, as of preprint submission, do not hold any public office.

We emphasize that the verification of (a)–(c) above required substantial manual effort, including direct contact with named individuals or their representatives. In one case (Output A), the named individual confirmed personal details consistent with the model output. We have no satisfying technical explanation for this and treat it as a curiosity warranting further investigation.

5.2 Possible Explanations

We consider several candidate explanations:

1. **Overfitting artifact.** The most parsimonious explanation is that LTGI at high α produces outputs that are technically out-of-distribution but appear meaningful due to the cognitive tendency to find patterns in noise. This explanation, however, does not account for the verifiable correctness of details in Output A.

2. **Data contamination.** The model may have been exposed during pretraining to information that was either inadvertently included or web-scraped. We performed extensive retrieval over public web sources and did not identify candidate sources for the specific details in question.

3. **Genuine extrapolation.** The model may, through some mechanism we do not understand, be performing inference over the temporal manifold that yields outputs that happen to correspond to actual future states of the world.

We do not endorse explanation (3). We note however that we have been unable to fully refute it through the means available to us, and that we consider this a serious limitation of our current methodology.

6 Discussion and Limitations

The findings reported in Section 5 pose methodological challenges we are not equipped to resolve. We initially viewed LTGI as a contribution to the interpretability literature on temporal representations. The anomalies we have documented suggest, at minimum, that the geometry of latent temporal manifolds in fine-tuned models is less well understood than the existing literature implies.

Several limitations warrant note:

Reproducibility. The anomalous outputs in Section 5 occurred at a low rate (estimated 0.4% of LTGI cycles at $\alpha = 4.0$) and have proven difficult to reproduce systematically. We have not been able to identify the specific conditions under which they arise.

Verification. Verifying the (non-)attestation of specific named entities, addresses, and dates in a corpus of 4.2B tokens is a non-trivial task. We cannot exclude the possibility that one or more anomalous outputs in fact derive from training data that we failed to identify.

Ethical considerations. The verification of Output A required contacting a third party regarding personal details that originated from a model. In retrospect, this was inadvisable. We strongly discourage further investigation along this axis without appropriate institutional review.

Author note. *This manuscript was withdrawn by the authors following internal review. We do not intend to pursue further research in this direction and we strongly discourage attempts to replicate the LTGI procedure as described. Inquiries should not be directed to the corresponding authors.*

7 Conclusion

We have introduced Layer-wise Temporal Gradient Inversion (LTGI) as a technique for manipulating latent temporal representations in fine-tuned language models. We have shown that LTGI produces measurable shifts in the temporal characteristics of generated text, and we have documented a class of anomalous outputs that we cannot fully explain. We recommend that this line of inquiry not be pursued.

References

Choi, S., Verma, A., & Tanaka, K. (2023). Time-coded representations in multilingual transformer models. *Transactions of the Association for Computational Linguistics*, 11, 1142–1161.

Demir, E., & Acar, B. (2023). Diachronic language modeling for low-resource Romance varieties. In *Proceedings of EACL 2023*, pages 2014–2027.

Gurnee, W., Bau, D., & Tegmark, M. (2023). Language models represent space and time. *arXiv preprint arXiv:2310.02207*.

Hernandez, E., Sharma, A., & Andreas, J. (2023). Inspecting and editing knowledge representations in language models. *arXiv preprint arXiv:2304.00740*.

Huang, L., Yu, W., Ma, W., et al. (2023). A survey on hallucination in large language models. *arXiv preprint arXiv:2311.05232*.

Kaplan, R., & Sørensen, M. (2022). Pretraining strategies for historical language modeling. In *Proceedings of LREC 2022*, pages 4012–4021.

Lampinen, A., Roy, N., & Saxe, A. (2024). Temporal probing of language models: A geometric analysis. *ICLR 2024*.

McKenna, N., Li, T., & Cheng, L. (2023). Sources of hallucination in language models. *Findings of EMNLP 2023*, 2391–2405.

Touvron, H., Martin, L., Stone, K., et al. (2023). LLaMA 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Turner, A., Thiergart, L., Udell, D., et al. (2023). Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*.

Zou, A., Phan, L., Chen, S., et al. (2023). Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*.