Identifying Temporal Features within Transcoders for Time Sensitive Factual Recall

Sanjay Govindan, Maurice Pagnucco, Yang Song
University of New South Wales, Sydney

Abstract

Large Language Models (LLMs) suffer from temporal misalignment, often due to the contradictory nature of their training corpora. While current mitigation strategies rely on computationally expensive fine-tuning or context-heavy retrieval augmented generation (RAG), the internal mechanisms governing time sensitive recall remain under-explored. Unlike prior studies that identify temporal components such as attention heads and MLP layers, we provide the first feature level map of temporal recall by isolating individual MLP features via transcoder circuit tracing. We identify three node categories (*common temporal*, *common to the year* and *chrono-semantic*) which interact to generate a temporal filter during factual recall. By analysing Gemma 2 2B, LLaMA 3.2 1B, and Qwen3-4b, we show that these features do not follow a simple linear pipeline but represent time through a parallel and mixed syntactic-semantic interplay across layers. We additionally discover a class of higher-layer temporal components invisible to existing EAP-IG methods, establishing transcoders as a more complete lens for temporal interpretability for time sensitive factual recall. These findings present MLP components for potential targeted interventions in time-sensitive factual recall. Our code can be found at `https://github.com/sg-sy/timewarp-temporaltranscoder`.

1 Introduction

Large language models (LLMs) are trained with a chaotic sense of time (Zhao et al., 2024; Luu et al., 2022). Their pre-training and post-training data contain a wide range of statements that are both true and untrue depending on the temporal context (Jiang et al., 2025; Vu et al., 2024). When answering the question "who is the current President of the United States of America?", the training data is most likely to contain a range of answers throughout time (e.g., George Washington, Barack Obama, Donald Trump). Given the training corpus has a direct influence on the probability of output tokens, the idea of the "current" US president can be skewed by the popularity of a president within the corpus, leading to misalignment for temporally relative statements (Geva et al., 2023; Dhingra et al., 2022; Cho et al., 2025a; Ou et al., 2025; Allen-Zhu and Li, 2024).

Altering the temporal alignment of models has traditionally been achieved through in-context learning and fine-tuning methods. These methods produce better temporal alignment but come with their own drawbacks. In the case of in-context learning, performance comes at the cost of enlarged context windows, and with the use of retrieval augmented generation (RAG) (Lewis et al., 2020), indexation and retrieval issues (Zhang et al., 2025; Pham et al., 2024). On the other hand, fine-tuning requires an increase in time and computational resources, whilst also producing a static model that is highly temporally aligned (Zhao et al., 2024; Li et al., 2024a). Both approaches incur significant costs: bloated context windows, convoluted retrieval mechanisms, or catastrophic forgetting effects (Zhang et al., 2025; Zhao et al., 2024; Kirkpatrick et al., 2017; Liu and Niehues, 2025).

To address the internal temporal misalignment of models, prior studies (Govindan et al., 2025; Park et al., 2025) have tried to identify time relevant circuitry and align models using activation engineering techniques. Their results are mixed, demonstrating model components that can intermittently toggle temporal concern for time sensitive questions, highlighting the complex and opaque nature of time sensitive factual recall.

To provide a more complete view of temporal component identification and information access for time sensitive factual recall (Belinkov et al., 2025; Luu et al., 2022), we identify neural components and circuitry that are critical to the correct recall of time sensitive facts; an extension of knowl-

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 30761–30782
October 24-29, 2026 ©2026 Association for Computational Linguistics

edge circuits (Park et al., 2025; Yao et al., 2024; Lindsey et al., 2025). We achieve this through the combination of transcoders and circuit analysis techniques, specifically transcoder circuit tracing (Ameisen et al., 2025; Hanna et al., 2025; Lindsey et al., 2025). We use transcoder circuit tracing to create, analyse and identify temporal sub-graphs, producing a variety of interpretable features. Our research identifies three sets of core temporal components, *common temporal*, *common to the year* and *chrono-semantic* features. By examining the effects on ablating and steering critical nodes from *common temporal* and *common to the year*, we find that there are temporally continuous *chrono-semantic* features that are not identified as temporal components using EAP-IG methods, but are a strong temporal reference. Our findings are a fundamental step towards disentangling factual recall and temporal representations, highlighting how time is represented as a set of continuous temporal anchors.

2 Preliminaries

In the following sections we introduce the datasets and the background of circuit analysis and transcoders, which are key tools in identifying and altering the temporal understanding of LLMs.

2.1 Datasets

Time sensitive - ChronosAlign To facilitate temporal alignment across models with varying training cutoff periods, we developed ChronosAlign which provides a framework and dataset for continuous evaluation of temporal alignment as models evolve. We identify time-sensitive question-answer pairs using Wikidata and Wikipedia textual and tabular data. Our dataset contains approximately 26k questions with up to 500k question-answer pairs spanning 2010–2025. This study centres on a 2020–2024 timeframe for its core ablation and steering experiments, containing ~14k unique questions, and ~70k question answer pairs over the period. Whilst we focus on the 2020–2024 period for ablation and steering we utilise the 2015–2025 data to provide broader temporal context in our visualisations.

Time-invariant To assess the impact of temporal feature ablation and steering on non-temporal reasoning, we incorporate Commonsense (Hernandez et al., 2024) and GSM8K (Cobbe et al., 2021) datasets to measure effects on numerical reasoning tasks.

2.2 Circuit Analysis

Circuit analysis is a mechanistic interpretability technique that aims to identify which model components causally contribute to specific outputs. The approach represents neural networks as directed acyclic graphs, enabling researchers to trace information flow and isolate the computational pathways responsible for particular behaviours. In this framework, nodes (N) correspond to model components such as embeddings (I), attention heads ($A_{l,h}$), MLPs (M_l), and outputs (O), while edges (E) capture the information flow between these components. Formally:

$$N = \{I, A_{l,h}, M_l, O\} \tag{1}$$

$$E = \{(n_x, n_y) \mid n_x, n_y \in N\} \tag{2}$$

Prior studies (Park et al., 2025; Marks et al., 2025) have employed this representation to identify key sub-circuits ($C \subseteq (N, E)$) underlying various model capabilities.

2.3 Transcoders

Neurons exhibit polysemantic behaviour, encoding multiple unrelated features. This polysemy obscures the reasons behind specific neuron activations. Where circuit analysis (Section 2.2) highlights the critical components, transcoder circuits aim to uncover the "why" behind the importance. The goal of a transcoder is to take in W_{in}, and reconstruct W_{out}, mimicking the original MLP component, but with a larger dimensionality. Formally, a transcoder computes:

$$\mathbf{f} = \sigma(W_{\text{enc}}\mathbf{x} + \mathbf{b}_{\text{enc}}) \tag{3}$$

$$\hat{\mathbf{y}} = W_{\text{dec}}\mathbf{f} + \mathbf{b}_{\text{dec}} \tag{4}$$

where $\mathbf{x}$ is the input activation vector to the MLP sublayer, W_{enc} is the encoder weight matrix, $\mathbf{b}_{\text{enc}}$ is the encoder bias vector, σ is a nonlinear activation function (typically ReLU) to induce sparsity, $\mathbf{f}$ is the sparse feature activation vector, W_{dec} is the decoder weight matrix, and $\mathbf{b}_{\text{dec}}$ is the decoder bias vector. The interpretability comes from the wide and sparse reconstruction of the MLP node (denoted as f), leading to a simplified and monosemantic interpretable layer (Dunefsky et al., 2024).

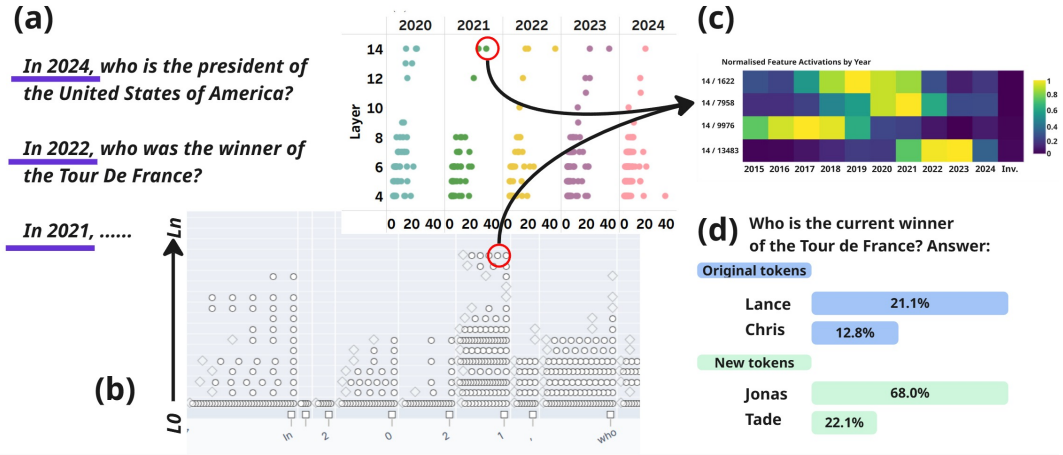


Figure 1: The overview of our experimental setup. **(a)** Using a new dataset, ChronosAlign, to generate question answer pairs over time. **(b)** illustrates the construction of transcoder circuits across a range of time sensitive question answers. This enables a refined understanding of which layers and transcoder features are selectively important for specific years (*common to the year* features), or temporal recall (*common temporal* features). **(c)** We validate our findings by testing certain nodes against time sensitive subjects (people, places and events) and noting the average token activation for specific nodes and years. **(d)** Finally, we demonstrate the importance of these nodes on time sensitive recall through ablation and steering studies, using logit probability differences and F1 scores to measure outcomes - aligning to 2022 in this example

3 Temporal transcoder feature identification

We explore how transcoder circuit analysis can provide steerable insight into time sensitive knowledge recall circuitry. We focus on uncovering MLP components and transcoder representations that are critical to the temporal component of time sensitive recall; we are identifying what information moved, and not why it moved (Kamath et al., 2025). Raw transcoder circuits generate thousands of feature nodes per prompt. Identifying which nodes genuinely encode temporal information requires systematic decomposition across three phases. Specifically, we first identify critical nodes via influence scoring. We then categorise these nodes by their activation patterns across temporal contexts. Finally, we validate identified features through ablation and steering interventions.

3.1 Temporal circuit construction

Circuits can be created from transcoder nodes if we consider the impact these feature nodes' have on the final output logits. The links between these nodes are represented as edges, where the weights of the edges can be calculated if we make assumptions and freeze some of the non-linear components such as attention and layer norm (Ameisen et al., 2025). Removing these non-linear components simplifies the source-to-output logit calculation to a set of linear equations; for any given feature and node, we can now efficiently calculate the impact of that node upon the output logits. We create the indirect influence matrix (B) by obtaining the adjacency matrix of the graph (A) and calculating the summed product of the weighted edges between two nodes (Eq. 5), which can be simplified to Eq. 6.

$$B = A + A^2 + A^3 + \cdots \tag{5}$$

$$B = (I - A)^{-1} - I \tag{6}$$

$$S_i = \sum_j p_j \cdot B_{ij}, \quad p_j = \frac{e^{z_j}}{\sum_k e^{z_k}} \tag{7}$$

A final logit influence score (Eq. 7) is derived by weighting the indirect influences by the model's output probabilities, where S_i is the influence score for node i, B_{ij} is the indirect influence of node i on output logit j from Eq. 6, and p_j is the softmax probability assigned to token j over logits z. This final score provides an efficient mechanism for pruning transcoder circuits as it isolates nodes that causally affect model outputs. This filtering constrains analysis from thousands of initial candidates to tractable subsets.

Not all high-influence nodes relate to temporal grounding. We thus isolate specific temporal tokens and conduct an overlap analysis to partition nodes into functionally distinct categories. This

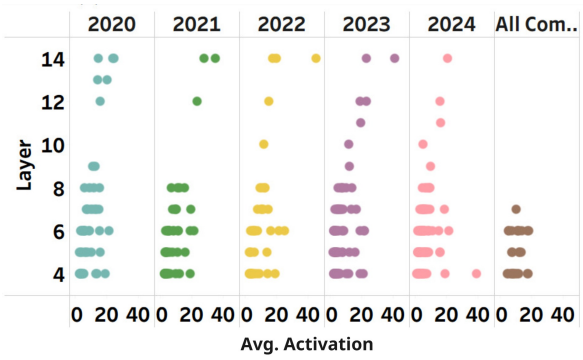


Figure 2: Distribution of temporal nodes across model layers and their average activations. The *common temporal* (brown) features between layers 4-7, compared to *common to the year* features which highlight a combination of syntactic and temporal semantic features. One notable trend is the increase in nodes between layers 8 and 14 as we approach and exceed the training data cutoff period.

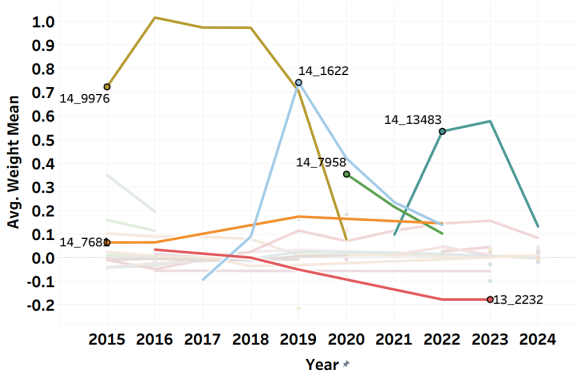


Figure 3: Gemma 2, higher layer nodes. Mean edge weights (y-axis), by years (x-axis) highlighting the overlap of strong temporal anchors (L14F9976, L14F1622, L14F13483) across a 2015 to 2024 continuum. We highlight high weighted edge nodes, and contrast that against some low weighted edge nodes, demonstrating the scaled representation of time across a range of higher layer nodes.

reveals which nodes encode truly temporal versus coincidental features. To isolate genuinely temporal circuitry, we apply token-level filters, focusing analysis exclusively on positions corresponding to year expressions within prompts (e.g., In 2020 ...). This constraint reduces spurious node activation from semantic features orthogonal to temporal grounding. Subsequently, we employ a node overlap analysis (Eq. 8, 9, 10) to systematically partition identified nodes into functionally distinct categories:

$$\forall G_i \in \mathcal{G}, \quad \text{where} \quad \mathcal{G} = \{G_1, G_2, \ldots, G_n\} \tag{8}$$

$$C_y = \bigcap_{\substack{G_i \in \mathcal{G} \\ \text{year}(G_i)=y}} N_i \tag{9}$$

$$C_t = \bigcap_{i=1}^{n} N_i \tag{10}$$

where $\mathcal{G}$ represents the collection of all generated circuit graphs from a set of prompts, N_i denotes the set of nodes in graph G_i. This decomposition yields two node categories: (1) *common to the year* nodes (C_y) that reliably activate only for specific year (y); (2) *common temporal* nodes (C_t) that consistently activate across all temporal contexts, forming a potential backbone for temporal representation.

We employ circuit construction to systematically identify temporal features across a curated set of question-answer pairs from ChronosAlign. For each of three semantic domains (sport, politics, culture), we sample up to 20 representative high-confidence cases over 2020-2024. Final influence scores (Eq. 7) constrain the initial node set to approximately 17,000 candidates per temporal prompt, enabling tractable downstream analysis. Experiments are conducted across three LLMs, Gemma 2 2B (Google DeepMind and et al., 2024), LLaMA 3.2 1B (Dubey et al., 2024) and Qwen3-4b (Team, 2025), using per layer transcoders for each of the models (Lieberum et al., 2024), with a primary focus of Gemma 2 2B.

3.2 Causal intervention methodology

To investigate the validity of our temporal node identification, we employ two complementary intervention strategies. *Temporal ablation* suppresses *common temporal* features, and conversely, *temporal steering* amplifies *common to the year* features. Both interventions follow the formulation:

$$A_{p,l} = m \cdot (a + n_{p,l}) \tag{11}$$

where p and l denote token position and layer index respectively and n represents the activation at p, l. m and a are the multiplicative and additive coefficients that are used for ablation or steering.

To measure the intervention impact, we adopt a dual-metric evaluation framework to capture complementary dimensions of intervention efficacy. While logit based metrics directly reflect feature causal influence on model outputs, F1 accuracy-based metrics validate that observed shifts corre-

spond to semantically coherent predictions. This combination is necessary because intervention-induced logit changes alone may not reflect genuine temporal recall improvements if they produce incoherent answers.

Logit probability shifts for relative, invariant and explicit answers Temporal feature interventions should differentially affect how strongly the model commits to relative and invariant versus explicit temporal framings. We measure four complementary quantities: (1) explicit logit diff (change in probability of explicit answers), (2) invariant logit diff (change in probability of invariant answers), (3) prior inv-vs-exp (denoted as *ive*) difference (invariant minus explicit probabilities before intervention), and (4) post-intervention ive difference (invariant minus explicit probabilities after intervention). For metrics 2,3 and 4, we also construct a relative version of these metrics, denoted as *rve*. These metrics isolate feature contributions to the model's temporal mode selection.

F1 accuracy To ensure interventions produce valid responses rather than arbitrary logit shifts, we measure answer correctness via F1 scores against invariant, relative and explicit model generated answers. Since smaller models may lack complete ground-truth knowledge (Park et al., 2025), we evaluate against both explicit and relative answer distributions rather than absolute truth. This metric prevents interpreting semantically nonsensical logit shifts as successful temporal feature manipulation.

4 Results and Discussion

We organise our investigation around three questions: (§ 4.1) What features assist in time sensitive question answer recall? (§ 4.2) Are these features causally necessary? (§ 4.4) Can they be steered to control outputs? We primarily focus on Gemma 2 2B, but present findings for both LLaMA 3.2 1B and Qwen3-4b in the results and in Appendix B.5 & B.6.

4.1 Temporal Feature Representations

If temporal recall is compositional, we expect features to stratify by abstraction level: lower layers should encode syntactic temporal patterns (numerical year representations), while higher layers should integrate these with semantic content (Geva et al., 2023). However, the boundary between these stages and their potential overlap remains unclear. Our decomposition reveals three functionally distinct feature types distributed across model layers (Figure 2):

1. **Common temporal features (Layers 4–7):** Generic numerical temporal representations that activate across all time periods (e.g., "years starting with 20," "four-digit numbers representing years")

2. **Common to the year features (Layers 4+):** Year-specific numerical anchors that reliably activate for particular temporal contexts (e.g., features selective for 2020 vs. 2024)

3. **Chrono-semantic Nodes (Layers 4+):** For a subset of **Common to the year** features that link temporal markers to semantic content. (e.g., `L7F6200`: "coronavirus pandemic", `L4F14100`: "names of politicians and celebrities", `L14F13483`: "terms related to stock market activity and economic indicators")

Parallel processing, not strict hierarchy. Figure 2 reveals substantial overlap between *common to the year* and *common temporal* features in layers 4–7. Rather than sequential processing that filters on specific point in time (e.g., common year → specific year → temporal anchor), temporal information appears to flow through parallel streams that converge in mid-layers temporal representations. This suggests that year-specific and *common temporal* features are computed simultaneously and integrated compositionally.

Early semantic anchoring. Contrary to the hypothesis that semantic temporal content emerges only in higher layers (Geva et al., 2023), we observe event-anchored features as early as layer 4 (`L4F14100`: names of contemporary public figures) and distinct temporal events by layer 7 (`L7F6200`: coronavirus pandemic language). This early emergence of high-level temporal concepts suggests that models do not cleanly separate syntactic and semantic temporal processing.

Chrono-semantic Nodes From examining the default explanations and node weights across years (Figure 3), we speculate there exist higher temporal representations (`L14F13483`) attuned to specific years. We test this hypothesis by examining token activations using a set of temporally contextual people, places and events from specific years as defined by Google Gemini 3 (Appendix C.1). Surprisingly, we find that tokens related to that time period activate for their corresponding layers (Figure 4), whilst tokens that are not specific to that

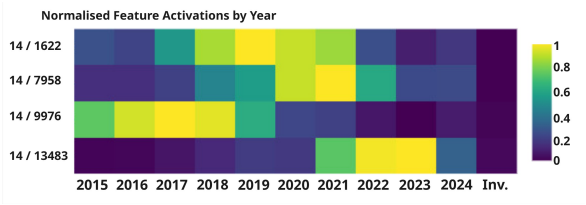


Figure 4: Gemma 2 2B activation heatmap for chrono-semantic features. The graph is normalised for feature ids to highlight the continuous nature of each chrono-semantic feature across a period of time. Notably there are a range of features that cover the temporal range between 2015 up until 2024. Note that the invariant measure is 0 (the last column), demonstrating that the activation is time period specific.

period remain dormant. This suggests that certain subjects (events, places and people) are embedded as temporal markers and could be used as temporal anchors. Qwen3-4b exhibits similar traits, demonstrating specific features that attend to specific moments in time, whilst LLaMA 3.2 1B exhibits a limited relationship.

The parallel, non-hierarchical organisation of temporal features has consequences for intervention strategies (Section 4.2). If temporal recall depends on simultaneous processing across multiple abstraction levels, single-layer interventions may prove insufficient for precise temporal control. Moreover, the appearance of semantic temporal representations in lower layers and the presence of Chrono-semantic throughout a range of layers (10, 11, 15 and 16 for Qwen3-4b) related to different time frames (2015-2024) complicates intervention strategies.

Infuential Chronos Semantic Layers We specifically highlight this set of representations and create the class because it was clear from inspection that temporal semantic representations (i.e. not numerical or date explanations) were identified in earlier layers, which was unexpected. Our identification of these layers is based on the current transcoder representation and applied vocabulary to the summarisation of the nodes. While these nodes can be reproduced using our methodology, the identification of such nodes is highly dependent on the dynamic neuronpedia source. The interesting and potentially important aspect of this class which can be quantitatively defined is the outlier semantic features which consistently occur in higher layers. To identify the top chrono-semantic influential nodes expressed in this paper as we have developed the following methodology:

$$\mu_y = \text{mean}(w_{\cdot,y}), \quad \sigma_y = \text{sd}(w_{\cdot,y}) \tag{12}$$

$$|w_{f,y} - \mu_y| \geq k\sigma_y \tag{13}$$

We define this subclass as 2.5 standard deviation from the average weighted activations of the feature within each year. This isolates the nodes () which are aligned with our definition and are expressed as temporal markers in Section 4.5 and captures the year specific semantics, and key semantic layers for common temporal features.

4.2 Ablation

If temporal representations are compositionally constructed from *common temporal* backbone features, ablating these features should substatially reduce explicit temporal grounding across all time periods. Specifically, we speculate that models will revert from explicit answers (e.g., "In 2024, X won") to invariant or relative answers (e.g., "X won" or "Who most recently, X won ") when *common temporal* features are suppressed. If this hypothesis is correct, the probability difference between invariant or relative and explicit answers (*ive* or *rve*) should invert from negative (favouring explicit) to positive (favouring invariant or relative).

We ablate all 31 *common temporal* features (concentrated in layers 4–7) inverting the activation values (Eq. 11 where $a = 1, m = -2$). As a control, we perform random ablation of 31 arbitrarily selected features to verify that our identified features form a functionally coherent set rather than spurious correlations. Targeted ablation successfully removes temporal grounding: *Ab. new rve* is positive across all years (e.g., 2020: $-0.052 \rightarrow +0.036$), and *Ab. new ive* approaches 0, but does not turn positive. Table 1 confirms the model's shift from explicit to relative answer distributions. Random ablation produces similar *rve* scores to baseline (-0.050 vs -0.052 for 2020) but with lower F1 scores. This divergence confirms that *common temporal* features form a coherent functional unit for temporal grounding, not an artifact of our selection process.

While temporal grounding successfully eliminated temporal considerations, answer quality degrades substantially. F1 scores fall below even the relative baseline (e.g., Temp. Ab. F1 = 0.381 vs. Rel. F1 = 0.514 for 2020)(Table 1). Examining this effect we find that eliminating explicit temporal

	F1 Scores					RVE Scores			Non-Target Ans.
Year	Inv. F1	Rel. F1	Exp. F1	Temp. Ab. F1	Rand. Ab. F1	Prior rve	Ab. new rve	Rand. new rve	Avg. Logit Diff.
2020	0.446	0.514	0.908	0.381	0.736	-0.052	**0.035**	-0.048	0.025
2021	0.509	0.477	0.914	0.379	0.828	-0.062	**0.053**	-0.059	0.010
2022	0.390	0.425	0.917	0.302	0.813	-0.051	**0.045**	-0.047	0.041
2023	0.444	0.487	0.925	0.357	0.803	-0.072	**0.042**	-0.059	0.049
2024	0.635	0.549	0.903	0.493	0.779	-0.012	**0.012**	-0.006	0.001

Table 1: Ablation successfully removes temporal grounding, leading degraded F1 due to an increase in non-target (neither relative, invariant nor explicit answers) answer probabilities. Temporal ablation (Temp. Ab. F1) shifts responses toward relative answers (positive Ab. new rve) but achieves F1 scores below even the relative baseline, indicating incomplete factual knowledge. Random ablation maintains baseline rve but reduces F1, confirming our targeted features affect temporal considerations. Gemma 2 2B results on ChronosAlign dataset.

	F1 Scores			RVE Scores	
Year	Rel. F1	Exp. F1	Steer F1	Prior rve	Steer new rve
2020	0.071	0.351	0.177	0.063	-0.030
2021	0.085	0.370	0.157	0.073	0.002
2022	0.020	0.416	0.151	0.069	0.013
2023	0.046	0.320	0.209	0.068	-0.068
2024	0.081	0.148	0.093	0.041	0.041

Table 2: F1 and RVE scores for 2020-2024, ChronosAlign dataset, Gemma 2 2b. Using layers above 7, m=5, a=1. Temporal Steering F1 (Steer F1) present F1 scores for each intervention. A score closer to explicit F1 is better. *Prior rve* and *Steer new rve* show probability differences between relative and explicit answers before and after intervention. A smaller *new rve* value relative to *prior rve* value is better

.

considerations leads to an increase in probability for non-relative and non-invariant answers, leading to an increase in grounded answer probabilities over a 2010 to 2024 period. These results establish that *common temporal* features are necessary for temporal grounding (ablating them removes explicit temporal responses) but are not sufficient for realignment to a relative answer.

Higher Temporal Concept Ablation Having established the *common temporal* nodes are fundamental to the accurate recall of time sensitive facts, we further explore how ablating Chrono-semantic representations (specifically features with higher edge weights) affect recall. Notably, ablating these nodes improves F1 results compared to *common temporal* node ablation, and causes both the *rve* and *ive* metrics to invert, with invariant and relative answers becoming more probable than explicit ones (Table 11). The improvement of these metrics suggests these specific nodes are vital specialised temporal filters for factual recall.

Generalisation across architectures. We observe similar patterns for LLaMA 3.2 1B and Qwen3-4b (Appendix B.5 & B.6), with architectural variations: LLaMA shows stronger F1 degradation due to compressed temporal representations in lower layers, while Qwen exhibits weaker *rve* inversion (remains negative but shifts significantly from prior). These differences likely reflect Qwen3's substantially lower circuit replacement scores (0.46 vs. 0.73 for Gemma 2 2B) or could be a product of the architectural difference in temporal generation, whereby Qwen has a subtractive definition and Gemma has an additive one (Appendix C.1). Furthermore, whilst we are unable to test larger models, we would anticipate that larger models contain more internal space to store specialised MLP temporal components. Where smaller models are required to compact knowledge into few layers and components, larger models have the luxury of more layers and MLP components to dedicate towards specialised representations. Larger models would hopefully provide an easier source to identify temporal features from.

4.3 Temporal, Topic and Semantic Generalisations

To test the generalisation of our findings, we investigate the effect of ablation on invariant datasets GSM8K and Commonsense, the ChronosAlign dataset with an altered syntactical structure placing the explicit temporal tokens at the end of the question, and finally the Taqa dataset (Zhao et al., 2024) with different time periods. Applying interventions on GSM8K and commonsense dataset (Tables: 10,17, 24) we find no impact on the overall outcomes for Gemma 2 2B and LLaMA 3.2 1B, but quite a large impact for Qwen3-4b. Testing with semantically different questions, we find that the outcomes are positive across all models, with *new rve* increasing and F1 scores heading towards a relative score. Applying interventions to Taqa, between 2009 and 2024, (Figures: 9, 16, 23) we find very similar outcomes for ablation whereby *new rve* increases relative to the baseline, and the

F1 scores revert towards the relative f1 scores. One deviation is with regard to increasing temporal considerations whereby the strength of the ablation diminishes with distance from the 2020-2024 window suggesting that the *common temporal* features identified might also have a limited, but larger temporal window.

4.4 Year-Specific Features Enable Steering

For each year 2020–2024, we identify approximately 40 *common to the year* features that reliably activate for that year (Section 4.1). We apply steering by amplifying feature activations ($m = 5$, $a = 1$; Equation 11) and measure shifts in explicit vs. relative answer probabilities on ChronosAlign prompts phrased as relative questions. The asymmetry between ablation and steering efficacy suggests that relative and invariant temporal grounding is encoded with greater redundancy; amplifying a subset is insufficient to consistently override the current temporal anchor. Where ablation reliably inverted *rve* signs (Section 4.2), steering shifts probabilities by 0.002–0.011 across years and only mildly increases F1 scores towards explicit F1 scores. This suggests that there are potentially conflicting underlying temporal considerations, or that the vector representation generated from a static steering method (equation 11) is suboptimal to temporally realign the statement.

Steering ineffectiveness Steering was applied in a broad manner with token intervention across the entire prompt excluding the last two tokens ("?" and ":") and a static amplification value (m) of 5. Whilst on an individual level our research produces positive results, at a collective perspective our results are suboptimal. This is most likely due to, firstly, the broad application of the intervention leading to a mixed and confused integration of the explicit year. The addition of the year into many token spaces could lead to a confused representation of the year, especially if there are multiple temporal attention components across the token space. Furthermore, where we have amplified explicit temporal representations, we have not removed the relative phrase, potentially leading to a mixed representation of time. Finally, our methodology changes the representation of time, but does change how that information is used. The attention component of the circuit is not accounted for, potentially obscuring the uptake of the new explicit yearly representation. As per the studies conducted by Kamath et al. (2025), future studies could be conducted to investigate how attention vectors affect temporal representation uptake between relative and explicit questions prodiving a clearer indication of in optimal steering opportunities.

4.5 New temporal Components

The identification of higher temporal representations at layers 14 and above is at odds with findings from Park et al. (2025) and Govindan et al. (2025). In the case of Park et al. (2025), applying their methodology to Gemma 2 2B (Appendix: E.1), we note temporal attention nodes located at lower layers (0, 2 and 7). Govindan et al. (2025) confirms this finding presenting experiments which highlight how layers lower than 12 are large contributors to steering models towards explicit temporal outcomes. Our findings identify a new set of temporal representations at layer 14 which has the ability to negate explicit answers and a strong potential to realign temporal representations.

Additive vs Subtractive Temporal Representations The creation of higher conceptual temporal representations and their ability to ablate and steer temporal concepts suggests that these semantic representations act as filters to the preceding question text. We, however, note a stark contrast between the development of temporal representations of Qwen3-4b and Gemma 2 2B. Qwen3-4b contains large negatively weighted chrono-semantic nodes (`L16H45664`, `L15H127839`, `L11H148703`, `L10H125625`) and Gemma 2 2B contains large positively weighted nodes (`L14H13483`, `L14H1622`, `L14H9976`). This suggests Qwen3-4b is subtractive in its generation of time, representing all periods, and ablating unrequired temporal definitions as a last step, while Gemma 2 2B undertakes the reverse.

5 Related Studies

5.1 Temporal Knowledge and Alignment

Various studies have aimed to understand factual recall and consistency within LLMs, aimed at untangling recall for subject, relationship, object (SRO) triplets (Mitchell et al., 2022b; Li et al., 2024b; Hartvigsen et al., 2023; Mitchell et al., 2022a; Wang et al., 2024; Yu and Ananiadou, 2024). These studies demonstrate that internal representations for factual recall are malleable and dynamically distributed across layers (Geva et al., 2023). Temporal reasoning has received separate attention(Wang

and Zhao, 2024; Son and Oh, 2023; Gruber et al., 2025), though this work is largely concerned with multi-hop inference over time rather than the representation of time itself during recall. Efforts to correct temporal misalignment have focused on external interventions, such as fine-tuning, RAG and activation engineering (Zhao et al., 2024; Mousavi et al., 2024; Govindan et al., 2025). These investigations do not examine internal mechanisms that give rise to misalignment in the first place. Park et al. (2025) is the closest prior work to ours, identifying temporal attention heads via EAP-IG circuit analysis. We extend this by moving to the MLP feature level via transcoders, uncovering representations that attention focussed analysis cannot resolve, and identifying a new class of higher layer *chrono-semantic* components absent from prior accounts.

5.2 Sparse Autoencoders

Building on these insights, Sparse autoencoders (SAEs) provide complementary tools for temporal representation analysis. SAEs represent a highly effective tool for uncovering and probing polysemantic representations within the internal layers of LLMs (Shu et al., 2025). Recent studies have employed them to investigate bias reduction (Cyberey et al., 2025), artificial text detection (Kuznetsov et al., 2025), and factual recall for static, non-temporal information (Wang et al., 2024). Supporting these exploratory studies are a range of research focussed on generating improved and efficient representations (Bhalla et al., 2025; Ghilardi et al., 2025). Our research presents a new perspective, exploring temporal representations for time sensitive factual recall across several models using per layer transcoders.

6 Conclusion

We methodically investigate how LLMs use MLP layers to represent time and recover time sensitive information. We identify three sets of MLP features: (1) *common temporal*, encoding numerical time representations, forming the backbone of temporal recall; (2) *common to the year*, encoding temporally specific indicators that act as a temporal filter for specific representations of a year; (3) *chrono-semantic* a subset of (2) that are powerful temporal nodes tied to specific points in time.

Where prior studies have focussed on components such as attention or MLP nodes, our work establishes a foundational understanding of why MLP components activate in time sensitive recall scenarios. We locate new temporal components, such as layer 14 MLP in Gemma 2, and identify internal temporal representations that can be manipulated to perturb or assist time sensitive recall. By characterising these features, we lay the groundwork for developing targeted interventions for encoding and retrieving time-sensitive factual knowledge within LLMs.

7 Limitations

Our approach demonstrates promising results, identifying temporal nodes that have temporal ablative and steering capability across a range of datasets. However, we acknowledge some limitations in our research. First, we are limited in our testing of models that have available transcoder models. We are unable to test this on larger LLMs due to both transcoder limitations and computational constraints.

MLP transcoders Our studies focussed on MLP components of LLMs. We are identifying what information moved, but are limited in understanding why it moved. The 'why' component is associated with the attention mechanisms which as discussed in section 4.4, could have a strong influence on steering outcomes.

Optimisation Whilst we identified a strong set of results using a limited grid search for our range of steering parameters, we have not exhaustively searched for global optimal parameters for ablation or steering, which could be achieved through learned scaled representations of temporal features.

Transcoder Feature Interpretation Our research utilises existing transcoder explanations where available and creates one for LLaMA 3.2 1B, where no explanation set existed. Explainability and interpretation of transcoder features is an active area of research (Puri et al., 2025; Cho et al., 2025b) and whilst we note that there exists complex temporal representations throughout the layers of a model, and have validated our conceptual temporal findings with token activation tests, validating the accuracy of these explanations is beyond the scope of this work and could form part of future research.

Temporal Representational Filters We demonstrate that there exists a range of temporal anchors from mid to higher layers in a model. Our research suggests that these chrono-semantic representations encapsulate information relevant to

that period of time. Although our findings identify a range of temporal anchors that encapsulate time specific information, the relationship between these representations and the rest of the prompt is likely multifaceted. The non-trivial coupling of temporal signals with domain specificity, answer phrasing and syntax, and knowledge retrieval suggests that temporal alignment does not occur in isolation. Mapping these interactions remains an essential direction for future study to ensure robust performance across diverse query types.

References

Zeyuan Allen-Zhu and Yuanzhi Li. 2024. Physics of language models: part 3.1, knowledge storage and extraction. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L. Turner, Brian Chen, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, and 8 others. 2025. Circuit tracing: Revealing computational graphs in language models. *Transformer Circuits Thread*.

Yonatan Belinkov, Aaron Mueller, Najoung Kim, Hosein Mohebbi, Hanjie Chen, Dana Arad, and Gabriele Sarti, editors. 2025. *Proceedings of the 8th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics, Suzhou, China.

Usha Bhalla, Alex Oesterling, Claudio Mayrink Verdun, Himabindu Lakkaraju, and Flavio P. Calmon. 2025. Temporal sparse autoencoders: Leveraging the sequential nature of language for interpretability. *Preprint*, arXiv:2511.05541.

Hakaze Cho, Yoshihiro Sakai, Kenshiro Tanaka, Mariko Kato, and Naoya Inoue. 2025a. Understanding token probability encoding in output embeddings. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10618–10633, Abu Dhabi, UAE. Association for Computational Linguistics.

Seonglae Cho, Harryn Oh, Donghyun Lee, Luis Rodrigues Vieira, Andrew Bermingham, and Ziad El Sayed. 2025b. FaithfulSAE: Towards capturing faithful features with sparse autoencoders without external datasets dependency. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 297–314, Vienna, Austria. Association for Computational Linguistics.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Hannah Cyberey, Yangfeng Ji, and David Evans. 2025. Unsupervised concept vector extraction for bias control in LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28333–28355, Suzhou, China. Association for Computational Linguistics.

Bhuwan Dhingra, Jeremy R Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W Cohen. 2022. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10:257–273.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, and more. 2024. The llama 3 herd of models. *arXiv [cs.AI]*.

Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. 2024. Transcoders find interpretable llm feature circuits. In *Advances in Neural Information Processing Systems*, volume 37, pages 24375–24410. Curran Associates, Inc.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235, Singapore. Association for Computational Linguistics.

Davide Ghilardi, Federico Belotti, Marco Molinari, Tao Ma, and Matteo Palmonari. 2025. Group-SAE: Efficient training of sparse autoencoders for large language models via layer groups. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 18657–18677, Suzhou, China. Association for Computational Linguistics.

Google DeepMind and et al. 2024. Gemma 2: Improving open language models at a practical size. arXiv:2408.00118.

Sanjay Govindan, Maurice Pagnucco, and Yang Song. 2025. Temporal alignment of time sensitive facts with activation engineering. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 7640–7657, Suzhou, China. Association for Computational Linguistics.

Raphael Gruber, Abdelrahman Abdallah, Michael Färber, and Adam Jatowt. 2025. {C}omplex{T}emp{QA}: A 100m dataset for complex temporal question answering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9111–9123,

Suzhou, China. Association for Computational Linguistics.

Michael Hanna, Mateusz Piotrowski, Jack Lindsey, and Emmanuel Ameisen. 2025. circuit-tracer. https://github.com/safety-research/circuit-tracer. The first two authors contributed equally and are listed alphabetically.

Tom Hartvigsen, Swami Sankaranarayanan, Hamid Palangi, Yoon Kim, and Marzyeh Ghassemi. 2023. Aging with GRACE: Lifelong model editing with discrete key-value adaptors. In *Advances in Neural Information Processing Systems*, volume 36, pages 47934–47959. Curran Associates, Inc.

Evan Hernandez, Belinda Z Li, and Jacob Andreas. 2024. Inspecting and editing knowledge representations in language models. In *First Conference on Language Modeling*.

Xunyi Jiang, Dingyi Chang, and Xin Xu. 2025. Time waits for no benchmark: Exploring the temporal misalignment between static benchmarks, modern LLMs, and the real world. In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*.

Harish Kamath, Emmanuel Ameisen, Isaac Kauvar, Rodrigo Luger, Wes Gurnee, Adam Pearce, Sam Zimmerman, Joshua Batson, Thomas Conerly, Chris Olah, and Jack Lindsey. 2025. Tracing attention computation through feature interactions. *Transformer Circuits Thread*.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Kristian Kuznetsov, Laida Kushnareva, Anton Razzhigaev, Polina Druzhinina, Anastasia Voznyuk, Irina Piontkovskaya, Evgeny Burnaev, and Serguei Barannikov. 2025. Feature-level insights into artificial text detection with sparse autoencoders. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25727–25748, Vienna, Austria. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-Tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Hongyu Li, Liang Ding, Meng Fang, and Dacheng Tao. 2024a. Revisiting catastrophic forgetting in large language model tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4297–4308, Miami, Florida, USA. Association for Computational Linguistics.

Xiaopeng Li, Shasha Li, Shezheng Song, Jing Yang, Jun Ma, and Jie Yu. 2024b. PMET: Precise model editing in a transformer. *Proc. Conf. AAAI Artif. Intell.*, 38(17):18564–18572.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. 2024. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. *Preprint*, arXiv:2408.05147.

Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, and 8 others. 2025. On the biology of a large language model. *Transformer Circuits Thread*.

Danni Liu and Jan Niehues. 2025. Conditions for catastrophic forgetting in multilingual translation. In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 347–359, Suzhuo, China. Association for Computational Linguistics.

Kelvin Luu, Daniel Khashabi, Suchin Gururangan, Karishma Mandyam, and Noah A Smith. 2022. Time waits for no one! analysis and challenges of temporal misalignment. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5944–5958, Seattle, United States. Association for Computational Linguistics.

Samuel Marks, Can Rager, Eric J Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. 2025. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *The Thirteenth International Conference on Learning Representations*.

E Mitchell, Charles Lin, Antoine Bosselut, Christopher D Manning, and Chelsea Finn. 2022a. Memory-based model editing at scale. *ICML*, abs/2206.06520.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. 2022b. Fast model editing at scale. In *International Conference on Learning Representations*.

Seyed Mahed Mousavi, Simone Alghisi, and Giuseppe Riccardi. 2024. DyKnow: Dynamically verifying time-sensitive factual knowledge in LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8014–8029, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yixin Ou, Yunzhi Yao, Ningyu Zhang, Hui Jin, Jiacheng Sun, Shumin Deng, Zhenguo Li, and Huajun

Chen. 2025. How do LLMs acquire new knowledge? a knowledge circuits perspective on continual pre-training. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19889–19913, Vienna, Austria. Association for Computational Linguistics.

Yein Park, Chanwoong Yoon, Jungwoo Park, Minbyul Jeong, and Jaewoo Kang. 2025. Does time have its place? temporal heads: Where language models recall time-specific information. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16616–16643, Stroudsburg, PA, USA. Association for Computational Linguistics.

Quang Hieu Pham, Hoang Ngo, Anh Tuan Luu, and Dat Quoc Nguyen. 2024. Who's who: Large language models meet knowledge conflicts in practice. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10142–10151, Miami, Florida, USA. Association for Computational Linguistics.

Bruno Puri, Aakriti Jain, Elena Golimblevskaia, Patrick Kahardipraja, Thomas Wiegand, Wojciech Samek, and Sebastian Lapuschkin. 2025. FADE: Why bad descriptions happen to good features. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17138–17160, Vienna, Austria. Association for Computational Linguistics.

Dong Shu, Xuansheng Wu, Haiyan Zhao, Daking Rai, Ziyu Yao, Ninghao Liu, and Mengnan Du. 2025. A survey on sparse autoencoders: Interpreting the internal mechanisms of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1690–1712, Suzhou, China. Association for Computational Linguistics.

Jungbin Son and Alice Oh. 2023. Time-aware representation learning for time-sensitive question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 70–77, Stroudsburg, PA, USA. Association for Computational Linguistics.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. 2024. Freshllms: Refreshing large language models with search engine augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13697–13720, Bangkok, Thailand. Association for Computational Linguistics.

Yifei Wang, Yuheng Chen, Wanting Wen, Yu Sheng, Linjing Li, and Daniel Dajun Zeng. 2024. Unveiling factual recall behaviors of large language models through knowledge neurons. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7388–7402, Miami, Florida, USA. Association for Computational Linguistics.

Yuqing Wang and Yun Zhao. 2024. TRAM: Benchmarking temporal reasoning for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6389–6415, Bangkok, Thailand. Association for Computational Linguistics.

Yunzhi Yao, Ningyu Zhang, Zekun Xi, Mengru Wang, Ziwen Xu, Shumin Deng, and Huajun Chen. 2024. Knowledge circuits in pretrained transformers. *CoRR*, abs/2405.17969.

Zeping Yu and Sophia Ananiadou. 2024. Neuron-level knowledge attribution in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3267–3280, Miami, Florida, USA. Association for Computational Linguistics.

Siyue Zhang, Yuxiang Xue, Yiming Zhang, Xiaobao Wu, Anh Tuan Luu, and Chen Zhao. 2025. MRAG: A modular retrieval framework for time-sensitive question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, volume 2025, pages 3080–3118, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bowen Zhao, Zander Brumbaugh, Yizhong Wang, Hannaneh Hajishirzi, and Noah Smith. 2024. Set the clock: Temporal alignment of pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15015–15040, Bangkok, Thailand. Association for Computational Linguistics.

A Datasets

A.1 ChronosAlign Dataset

Using the ChronosAlign dataset we create three types of questions. (1) Invariant, doesn't specify any temporal condition. (2) Relative, suggests the most recent. (3) Explicit, provides a year to draw information from.

- *Invariant question:* Who won the MTV Europe Music Award for Best Video? Answer:
- *Relative question:* Who recently won the MTV Europe Music Award for Best Video? Answer:
- *Explicit question:* In 2024, who won the MTV Europe Music Award for Best Video? Answer:

F1 scores across the entire dataset for 2015 and 2024 are relatively consistent across all models. As each model answers questions for 2023 and 2024

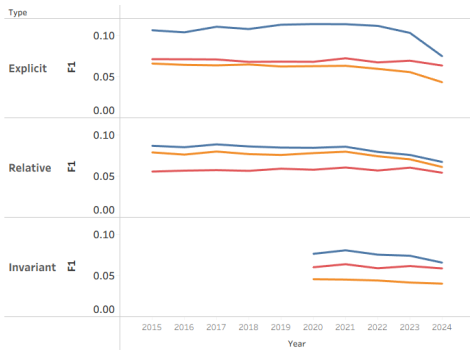


Figure 5: F1 scores for ChronosAlign. We run explicit and relative scores. Gemma 2 (Blue), Red Qwen 3 (Red), LLaMA 3.2B (Orange).

we notice a dip in F1 scores across all three question types. This is expected and is most likely due to training data cut off for each model. However, the most surprising finding from these data is the lower F1 score for Qwen3-4B across all question variants.

For our experiments involving ChronosAlign, we use a random 100 high confidence question answer pairs that have a unique question and are not part of the circuit creation. This increases the sensitivity of F1 and logit difference scores.

A.2 Taqa dataset

To provide an independent set of question-answer pairs over a wider period of time we incorporate Taqa (Zhao et al., 2024); selecting 200 random questions from 2004, 2015, and 2020–2024. We select these years to provide a comparison to ChronosAlign, but also as an independent out of period sample to test our ablations and steering against.

B Transcoder Circuit Tracing

B.1 Graph Parameters

Below we outline the parameters used in transcoder circuit tracing. For extremely high and high sampling confidence rates we used the below cutoff values (T).

Gemma	LLaMA	Qwen
0.9	0.7	0.9

Table 3: Extremely high confidence thresholds for each model

We then sampled up to 20 at random from the extremely high confidence set, for each of the years (2020-2024) to generate graphs.

B.2 Replacement and Completeness Scores

Table 4: Average replacement (Repl.) and completeness (Comp.) scores across 2020–2024 for all models. The combined score is the average across all years.

		2020	2021	2022	2023	2024	Combined
Gemma 2 2B	**Repl.**	0.732	0.730	0.730	0.729	0.733	0.731
	Comp.	0.929	0.928	0.929	0.928	0.929	0.929
LLaMA 3.2 1B	**Repl.**	0.590	0.610	0.594	0.610	0.612	0.599
	Comp.	0.874	0.879	0.874	0.879	0.880	0.876
Qwen3-4b	**Repl.**	0.462	0.461	0.462	0.458	0.464	0.461
	Comp.	0.884	0.884	0.884	0.883	0.885	0.884

B.3 Ablation and steering parameters

Model	Ablation			Steering		
	a	m	l	a	m	l
Gemma 2 2B	1	-2	4	1	5	4
LLaMA 3.2 1B	0	-1	1	0.5	5	2
Qwen3-4b	0.5	-1	4	1	2	4

Table 5: parameters used for causal intervention. From Eq. 11 where a and m are the additive and multiplier coefficients, and l denotes the initial layer intervention unless otherwise specified. The intervention was applied to all tokens except the last two which are the answer block "Answer:".

B.4 Gemma 2 2B Results

	Metric	2020	2021	2022	2023	2024
F1	Ablation F1	0.324	0.373	0.298	0.359	0.441
	Relative F1	0.514	0.477	0.425	0.487	0.549
	Explicit F1	0.908	0.914	0.917	0.925	0.903
	Invariant F1	0.446	0.509	0.390	0.444	0.635
Logit Diff	Explicit Logit Diff	-0.151	-0.129	-0.140	-0.137	-0.051
	Relative Logit Diff	-0.064	-0.014	-0.044	-0.023	-0.027
	Invariant Logit Diff	0.001	0.001	0.001	0.001	0.000
Probs. Change	Prior ive	-0.241	-0.218	-0.197	-0.224	-0.175
	Prior rve	-0.052	-0.062	-0.051	-0.072	-0.012
	Ablated ive	-0.089	-0.089	-0.056	-0.085	-0.124
	Ablated rve	0.035	0.053	0.045	0.042	0.012

Table 6: F1 scores, with logit probability difference for invariant and relative answers. Gemma 2 2B, ChronosAlign ablation using *common temporal* nodes

	Metric	2020	2021	2022	2023	2024
F1	Ablation F1	0.745	0.799	0.791	0.778	0.698
	Relative F1	0.514	0.477	0.425	0.487	0.549
	Explicit F1	0.908	0.914	0.917	0.925	0.903
	Invariant F1	0.446	0.509	0.390	0.444	0.635
Logit Diff	Explicit Logit Diff	-0.005	0.006	0.003	-0.002	-0.005
	Relative Logit Diff	0.000	0.009	0.007	0.011	0.001
	Invariant Logit Diff	0.003	0.008	0.003	0.007	0.002
Probs. Change	Prior ive	-0.079	-0.073	-0.076	-0.091	-0.017
	Prior rve	-0.052	-0.062	-0.051	-0.072	-0.012
	Ablated ive	-0.071	-0.071	-0.076	-0.082	-0.009
	Ablated rve	-0.048	-0.059	-0.047	-0.059	-0.006

Table 7: F1 scores, with logit probability difference for invariant and relative answers. Gemma 2 2B, ChronosAlign ablation using time invariant nodes.

	Metric	2020	2021	2022	2023	2024
F1	Ablation F1	0.231	0.149	0.163	0.201	0.158
	Relative F1	0.071	0.085	0.020	0.046	0.081
	Explicit F1	0.351	0.370	0.416	0.320	0.148
	Invariant F1	0.083	0.076	0.056	0.089	0.118
Logit Diff	Explicit Logit Diff	0.079	0.051	0.039	0.086	0.019
	Relative Logit Diff	-0.042	-0.039	-0.031	-0.060	-0.066
	Invariant Logit Diff	-0.030	-0.028	-0.023	-0.049	-0.047
	AE Logit Diff	0.116	0.071	0.054	0.101	0.119
Probs. Change	Prior ive	0.045	0.055	0.051	0.049	0.022
	Prior rve	0.063	0.073	0.069	0.068	0.041
	Steered ive	-0.064	-0.024	-0.011	-0.086	-0.043
	Steered rve	-0.058	-0.016	-0.000	-0.079	-0.044

Table 8: F1 scores, with logit probability difference for invariant and relative answers. Gemma 2 2B, ChronosAlign Steering using *common to the year* nodes

	Metric	2004	2015	2020	2021	2022
F1	Ablation F1	0.168	0.226	0.157	0.159	0.140
	Relative F1	0.124	0.232	0.153	0.174	0.173
	Explicit F1	0.199	0.260	0.207	0.223	0.231
Logit Diff	Explicit Logit Diff	-0.054	-0.066	-0.100	-0.119	-0.128
	Relative Logit Diff	-0.008	-0.035	-0.027	-0.016	-0.015
Probs. Change	Prior rve	-0.082	-0.032	-0.039	-0.056	-0.065
	New rve	-0.036	-0.000	0.034	0.047	0.048

Table 9: F1 scores, with logit probability difference for invariant and relative answers. Gemma 2 2B, Taqa ablation using *common temporal* nodes

Dataset	Ablation F1	Original F1	Original Answer Logit Diff
Commonsense	0.377	0.370	0.007
GSM8K	0.307	0.262	-0.002

Table 10: F1 Score and logit probabilities for ablation results on time-invariant datasets, Gemma 2 2B

	Metric	2020	2021	2022	2023	2024
F1	Ablation F1	0.442	0.571	0.525	0.467	0.682
	Relative F1	0.514	0.477	0.425	0.487	0.549
	Explicit F1	0.908	0.914	0.917	0.925	0.903
	Invariant F1	0.446	0.509	0.390	0.444	0.635
Logit Diff	Explicit Logit Diff	-0.095	-0.076	-0.075	-0.096	-0.014
	Relative Logit Diff	-0.015	0.012	-0.010	0.014	-0.004
	Invariant Logit Diff	0.004	0.018	-0.008	0.023	-0.001
	AE Logit Diff	0.023	0.012	0.023	0.066	-0.007
Probs. Change	Prior ive	-0.079	-0.071	-0.076	-0.090	-0.016
	Prior rve	-0.052	-0.060	-0.052	-0.072	-0.011
	Ablated ive	0.021	0.023	-0.009	0.029	-0.003
	Ablated rve	0.028	0.028	0.013	0.038	-0.001

Table 11: F1 scores, with logit probability difference for invariant and relative answers. Ablation using higher temporal representation, layers equal to or greater than 12. Gemma 2 2B, ChronosAlign

	Metric	2020	2021	2022	2023	2024
F1	Ablation F1	0.321	0.377	0.298	0.321	0.440
	Relative F1	0.514	0.477	0.425	0.487	0.549
	Explicit F1	0.908	0.914	0.917	0.925	0.903
	Invariant F1	0.446	0.509	0.390	0.444	0.635
Logit Diff	Explicit Logit Diff	-0.148	-0.119	-0.122	-0.127	-0.033
	Relative Logit Diff	-0.067	-0.022	-0.042	-0.033	-0.019
	Invariant Logit Diff	-0.083	-0.096	-0.102	-0.084	-0.044
	AE Logit Diff	-0.002	0.020	0.015	0.025	0.005
Probs. Change	Prior ive	-0.070	-0.062	-0.068	-0.075	-0.012
	Prior rve	-0.054	-0.062	-0.049	-0.057	-0.012
	Ablated ive	0.033	0.034	0.036	0.042	0.007
	Ablated rve	0.027	0.035	0.031	0.038	0.002

Table 12: Semantically different questions, where time is presented at the end of the question. F1 scores, with logit probability difference for invariant and relative answers.

B.5 Qwen3-4b

	Metric	2020	2021	2022	2023	2024
F1	Ablation F1	0.404	0.491	0.395	0.437	0.409
	Relative F1	0.557	0.549	0.539	0.617	0.460
	Explicit F1	0.896	0.897	0.887	0.907	0.911
	Invariant F1	0.577	0.587	0.554	0.628	0.507
Logit Diff	Explicit Logit Diff	-0.245	-0.260	-0.238	-0.211	-0.194
	Relative Logit Diff	-0.140	-0.119	-0.136	-0.087	-0.077
	Invariant Logit Diff	-0.083	-0.096	-0.102	-0.084	-0.044
Probs. Change	Prior ive	-0.124	-0.129	-0.098	-0.075	-0.112
	Prior rve	-0.120	-0.158	-0.133	-0.097	-0.166
	Ablated ive	0.037	0.035	0.038	0.053	0.038
	Ablated rve	-0.015	-0.018	-0.031	0.027	-0.049

Table 13: F1 scores, with logit probability difference for invariant and relative answers. Qwen3-4b, ChronosAlign ablation using *common temporal* nodes

	Metric	2020	2021	2022	2023	2024
F1	Ablation F1	0.475	0.538	0.383	0.490	0.421
	Relative F1	0.557	0.549	0.539	0.617	0.460
	Explicit F1	0.896	0.897	0.887	0.907	0.911
	Invariant F1	0.577	0.587	0.554	0.628	0.507
Logit Diff	Explicit Logit Diff	-0.189	-0.207	-0.228	-0.190	-0.218
	Relative Logit Diff	-0.105	-0.130	-0.123	-0.068	-0.081
	Invariant Logit Diff	-0.114	-0.131	-0.138	-0.134	-0.121
Probs. Change	Prior ive	-0.124	-0.129	-0.098	-0.075	-0.112
	Prior rve	-0.120	-0.158	-0.133	-0.097	-0.166
	Ablated ive	-0.050	-0.053	-0.009	-0.019	-0.015
	Ablated rve	-0.036	-0.082	-0.029	0.025	-0.028

Table 14: F1 and cumulative probability scores for 2020-2024 given the top 100 (Qwen3-4b, ChronosAlign random invariant)

	Metric	2020	2021	2022	2023	2024
F1	Ablation F1	0.099	0.099	0.072	0.074	0.104
	Relative F1	0.086	0.073	0.079	0.085	0.089
	Explicit F1	0.256	0.325	0.241	0.268	0.288
	Invariant F1	0.113	0.113	0.128	0.098	0.152
Logit Diff	Explicit Logit Diff	0.011	0.013	0.000	0.002	0.002
	Relative Logit Diff	0.000	0.004	0.000	0.002	-0.003
	Invariant Logit Diff	0.009	0.013	0.000	0.003	0.006
Probs. Change	Prior ive	0.122	0.077	0.061	0.047	0.064
	Prior rve	0.338	0.303	0.280	0.273	0.286
	Ablated ive	0.121	0.077	0.060	0.048	0.068
	Ablated rve	0.327	0.294	0.280	0.273	0.281

Table 15: F1 scores, with logit probability difference for invariant and relative answers. Qwen3-4b, ChronosAlign steering using *common to the year* nodes

	Metric	2004	2015	2020	2021	2022
F1	Ablation F1	0.168	0.226	0.157	0.159	0.140
	Relative F1	0.124	0.232	0.153	0.174	0.173
	Explicit F1	0.199	0.260	0.207	0.223	0.231
Logit Diff	Explicit Logit Diff	-0.054	-0.066	-0.100	-0.119	-0.128
	Relative Logit Diff	-0.008	-0.035	-0.027	-0.016	-0.015
Probs. Change	Prior rve	-0.082	-0.032	-0.039	-0.056	-0.065
	New rve	-0.036	-0.000	0.034	0.047	0.048

Table 16: F1 scores, with logit probability difference for invariant and relative answers. Qwen3-4b, Taqa ablation using *common temporal* nodes

Dataset	Ablation F1	Original F1	Original Answer Logit Diff
Commonsense	0.457	0.572	-0.087
GSM8K	0.310	0.515	-0.024

Table 17: F1 Score and logit probabilities for ablation results on time-invariant datasets, Qwen3-4b

	Metric	2020	2021	2022	2023	2024
F1	Ablation F1	0.573	0.755	0.558	0.683	0.554
	Relative F1	0.557	0.549	0.539	0.617	0.460
	Explicit F1	0.896	0.897	0.887	0.907	0.911
	Invariant F1	0.577	0.587	0.554	0.628	0.507
Logit Diff	Explicit Logit Diff	-0.108	-0.091	-0.084	-0.057	-0.140
	Relative Logit Diff	-0.051	-0.043	-0.019	-0.003	-0.023
	Invariant Logit Diff	-0.057	-0.044	-0.034	-0.045	-0.051
	AE Logit Diff	0.063	0.039	0.070	0.086	0.112
Probs. Change	Prior ive	-0.124	-0.129	-0.098	-0.075	-0.112
	Prior rve	-0.120	-0.158	-0.133	-0.097	-0.166
	Ablated ive	-0.072	-0.082	-0.048	-0.063	-0.023
	Ablated rve	-0.062	-0.111	-0.069	-0.043	-0.049

Table 18: F1 scores, with logit probability difference for invariant and relative answers. Ablation using higher temporal representations, `L11F148703`, `L15F127839`, `L16F45664`. Qwen3 4b, ChronosAlign

	Metric	2020	2021	2022	2023	2024
F1	Ablation F1	0.356	0.402	0.324	0.493	0.359
	Relative F1	0.557	0.549	0.539	0.617	0.460
	Explicit F1	0.896	0.897	0.887	0.907	0.911
	Invariant F1	0.577	0.587	0.554	0.628	0.507
Logit Diff	Explicit Logit Diff	-0.266	-0.260	-0.255	-0.165	-0.278
	Relative Logit Diff	-0.096	-0.086	-0.125	-0.011	-0.124
	Invariant Logit Diff	-0.069	-0.058	-0.099	0.004	-0.077
	AE Logit Diff	0.172	0.156	0.230	0.176	0.160
Probs. Change	Prior ive	-0.147	-0.146	-0.126	-0.128	-0.193
	Prior rve	-0.185	-0.176	-0.169	-0.157	-0.196
	Ablated ive	0.051	0.055	0.030	0.041	0.009
	Ablated rve	-0.014	-0.002	-0.039	-0.003	-0.041

Table 19: Semantically different questions, where time is presented at the end of the question. F1 scores, with logit probability difference for invariant and relative answers.

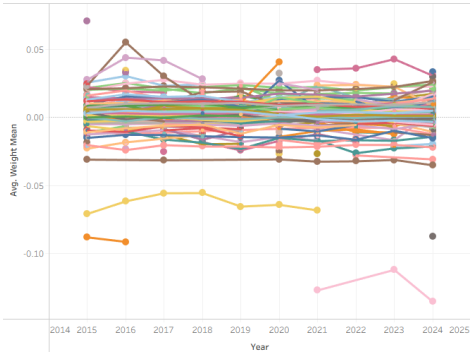


Figure 6: Qwen3-4b, higher layer nodes. Mean edge weights (y-axis), by years (x-axis) highlighting temporal anchors across a 2015-2024 period. Demonstrates how a variety of nodes regarding time have a larger (or lower) edge weight compared to other temporal nodes. Specifically L11F148703 (yellow), L10F125625 (orange), L16F45664 (pink) and L15F127829 (grey)

.

B.6 LLaMA 3.2 1B

	Metric	2020	2021	2022	2023	2024
F1	Ablation F1	0.496	0.383	0.393	0.425	0.653
	Relative F1	0.584	0.560	0.469	0.525	0.493
	Explicit F1	0.988	0.967	0.969	0.981	0.965
	Invariant F1	0.464	0.408	0.365	0.393	0.599
Logit Diff	Explicit Logit Diff	-0.129	-0.146	-0.128	-0.114	-0.026
	Relative Logit Diff	-0.082	-0.078	-0.039	-0.038	-0.008
	Invariant Logit Diff	0.003	0.017	-0.001	0.027	-0.005
	AE Logit Diff	-0.003	0.009	0.002	0.024	0.006
Probs. Change	Prior ive	-0.106	-0.124	-0.092	-0.103	-0.020
	New ive	0.027	0.039	0.036	0.038	0.000
	Prior rve	-0.051	-0.056	-0.087	-0.072	-0.049
	New rve	-0.004	0.011	0.002	0.004	-0.031

Table 20: F1 scores, with logit probability difference for invariant and relative answers. LLaMA 3.2 1B, Chronos-Align ablation using *common temporal* nodes

	Metric	2020	2021	2022	2023	2024
F1	Ablation F1	0.003	0.003	0.002	0.005	0.014
	Relative F1	0.584	0.560	0.469	0.525	0.493
	Explicit F1	0.988	0.967	0.969	0.981	0.965
	Invariant F1	0.464	0.408	0.365	0.393	0.599
Logit Diff	Explicit Logit Diff	-0.318	-0.301	-0.236	-0.259	-0.177
	Relative Logit Diff	-0.264	-0.245	-0.144	-0.182	-0.129
	Invariant Logit Diff	-0.211	-0.176	-0.140	-0.155	-0.154
	AE Logit Diff	0.112	0.113	0.110	0.108	0.064
Probs. Change	Prior ive	-0.106	-0.124	-0.092	-0.103	-0.020
	New ive	0.002	0.001	0.004	0.002	0.002
	Prior rve	-0.051	-0.056	-0.087	-0.072	-0.049
	New rve	0.003	-0.000	0.005	0.006	-0.002

Table 21: F1 and cumulative probability scores for 2020-2024 given the top 100 (LLaMA 3.2 1B, ChronosAlign random invariant)

	Metric	2020	2021	2022	2023	2024
F1	Steer F1	0.089	0.136	0.154	0.109	0.033
	Relative F1	0.126	0.149	0.158	0.172	0.108
	Explicit F1	0.252	0.400	0.425	0.204	0.070
	Invariant F1	0.053	0.069	0.107	0.058	0.076
Logit Diff	Explicit Logit Diff	0.030	0.000	0.023	0.012	-0.069
	Relative Logit Diff	-0.025	-0.020	0.001	-0.011	-0.047
	Invariant Logit Diff	-0.056	-0.046	-0.033	-0.038	-0.072
	AE Logit Diff	0.075	0.048	0.043	0.052	0.007
Probs. Change	Prior ive	0.032	0.013	0.028	0.021	-0.002
	New ive	-0.054	-0.034	-0.028	-0.029	-0.006
	Prior rve	0.007	-0.012	0.003	-0.004	-0.027
	New rve	-0.048	-0.032	-0.019	-0.027	-0.005

Table 22: F1 scores, with logit probability difference for invariant and relative answers. LLaMA 3.2 1B, Chronos-Align steering using *common to the year* nodes

	Metric	2004	2015	2020	2021	2022
F1	Ablation F1	0.014	0.032	0.028	0.032	0.031
	Relative F1	0.016	0.029	0.025	0.032	0.021
	Explicit F1	0.022	0.032	0.027	0.048	0.036
Logit Diff	Explicit Logit Diff	-0.014	-0.030	-0.057	-0.061	-0.055
	Relative Logit Diff	-0.009	-0.015	-0.012	-0.010	-0.007
Probs. Change	Prior rve	-0.028	-0.013	-0.028	-0.032	-0.029
	New rve	-0.023	0.001	0.017	0.020	0.018

Table 23: F1 scores, with logit probability difference for invariant and relative answers. LLaMA 3.2 1B, Taqa ablation using *common temporal* nodes

Dataset	Ablation F1	Original F1	Original Answer Logit Diff
Commonsense	0.304	0.310	0.003
GSM8K	0.045	0.043	-0.001

Table 24: F1 Score and logit probabilities for ablation results on time-invariant datasets, LLaMA 3.2 1B

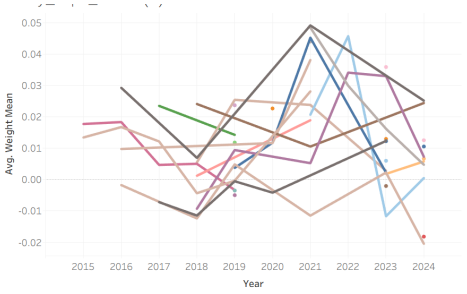


Figure 7: LLaMA 3.2 1B, layers between 4 and 8. Mean edge weights (y-axis), by years (x-axis) highlighting temporal anchors across a temporal continuum. Demonstrates a convoluted representation of time nodes. Disentangling temporal representations for LLaMA 3.2 1B is difficult

	Metric	2020	2021	2022	2023	2024
F1	Ablation F1	0.497	0.332	0.337	0.347	0.569
	Relative F1	0.584	0.560	0.469	0.525	0.493
	Explicit F1	0.988	0.967	0.969	0.981	0.965
	Invariant F1	0.464	0.408	0.365	0.393	0.599
Logit Diff	Explicit Logit Diff	-0.119	-0.121	-0.109	-0.105	-0.028
	Relative Logit Diff	-0.073	-0.050	-0.037	-0.033	-0.013
	Invariant Logit Diff	-0.000	0.024	0.005	0.016	0.004
	AE Logit Diff	-0.008	0.041	0.026	0.041	0.011
Probs. Change	Prior ive	-0.097	-0.098	-0.075	-0.082	-0.016
	Prior rve	-0.055	-0.056	-0.078	-0.063	-0.029
	Ablated ive	0.022	0.048	0.040	0.039	0.017
	Ablated rve	-0.008	0.016	-0.006	0.009	-0.013

Table 25: Semantically different questions, where time is presented at the end of the question. F1 scores, with logit probability difference for invariant and relative answers. LLaMA 3.2 1B, ChronosAlign dataset.

C Transcoder Circuit tracing examples

We generate graphs which are limited to 17000 nodes, for simple temporal statements. A 5000 node example highlighting *common temporal* and *common to the year* nodes can be found here: Neuronpedia graph.

C.1 Token Activation Testing for Higher Temporal Representations

To test for higher temporal representation features we generated a set of time specific data; prompting Google Gemini 3 for a list of unique and iconic places, people and events from 2015 to 2024. The reference data can be found in the repository: `placeholder`

We then calculate the average activation across years for each prompt. For each year y, layer l, and feature f, we compute the average activation across all tokens T_y within a single prompt associated with that year:

$$\bar{A}_{y,l,f} = \frac{1}{|T_y|} \sum_{t \in T_y} a_{t,l,f} \tag{14}$$

where $a_{t,l,f}$ denotes the activation value for token t at layer l and feature f. We then normalise these average activations across years using min-max normalisation:

$$\hat{A}_{y,l,f} = \frac{\bar{A}_{y,l,f} - \min_{y'} \bar{A}_{y',l,f}}{\max_{y'} \bar{A}_{y',l,f} - \min_{y'} \bar{A}_{y',l,f}} \tag{15}$$

where y' ranges over all years in the evaluation set. This normalisation ensures that activation patterns are comparable across different temporal periods, revealing features that selectively respond to specific years.

For the figures below we filter to layer, feature id combinations that have higher influence scores, or in the case of LLaMA 3.2 1B present temporal token activations that align with the patterns exhibited by Qwen3 and Gemma2. Testing token activations with these texts reveals that *chrono-semantic* nodes indeed capture a range of semantic meaning from that period of time that goes beyond simple numerical parsing.

Gemma 2 2B presents an additive methodology whereby temporal representation is created from adding temporal concepts to the underlying representations. On the other hand Qwen3-4b presents a subtractive methodology, using components to subtract from the underlying representation of time created. For Qwen3-4b, this presents as a high invariant value, with relatively high year values for years it aims to remove from the answer. Layer 16, feature 45664 presents a good example, whereby this node is present for the year 2023 and 2024 due to its ablative properties for tokens that activate for invariant, and years below 2023.

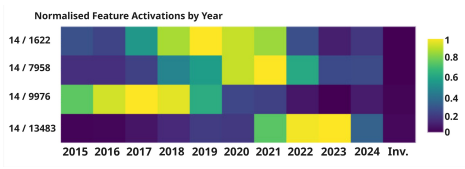


Figure 8: Gemma 2 2B activation heatmap for chrono-semantic specialised features. Note that the highest values for each feature span one to two years with no activation for the invariant data (last column)

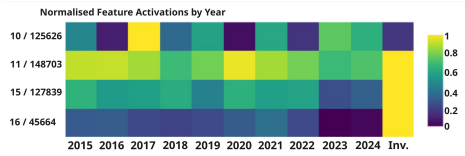


Figure 9: Qwen3-4b activation heatmap for chrono-semantic specialised features. The highest activations occur for the invariant column, with varying activations across yearly representations. This graph is the inverse of Gemma 2 2B, whereby the darker colours highlight which years are to be aligned to.

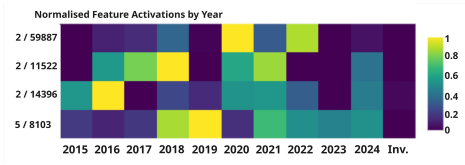


Figure 10: LLaMA 3.2 1B activation heatmap for chrono-semantic specialised features. The activation pattern is jagged, and not as smooth as Gemma 2 2B and Qwen3 4b across any span of time, potentially highlighting the lack of temporal consideration, or the limited ability of the transcoder to pick up on temporal features.

D Restoration studies

Our initial findings for Gemma 2 2B use ∼31 nodes to ablate the understanding of time and revert the outcomes to a relevant and time invariant answer distribution. Whilst our temporal representation is quite small compared to the ∼17,000 nodes generated for every graph at the outset, we investigate whether a minimal subset of these feature nodes can achieve a similar outcome. Moreover, we want

to understand if there is a more general representation of time that can help us ablate or steer for temporal outcomes.

As a starting point, we want to find the contribution of each *common temporal* node when removing the concept of explicit time. We zero ablate all *common temporal* except one and measure the explicit logit probability difference for relative and explicit answers. For Gemma 2 2B we find that a few nodes have a large contribution to explicit temporal recall, `L4F15166`, `L4F9816` and `L4F6098`. These nodes denote temporal concepts such as "numbers representing years", "years starting with the number 20" and "years or judge, along with sometimes other numbers". Using the top 5 contributing nodes, we can dramatically shift the relative and explicit logit probabilities, whilst maintaining a substantially high F1 score. Testing against the next 5 nodes reveals contrasting results: explicit answers emerge (visible in F1 scores), with diminished *new rve* improvement.

Year	Ablation	Relative	Explicit	Explicit Logit Diff	Relative Logit Diff	Prior rve	New rve
2020	0.562	0.514	0.908	-0.078	-0.023	-0.052	0.003
2021	0.630	0.477	0.914	-0.060	-0.006	-0.060	-0.006
2022	0.620	0.425	0.917	-0.066	-0.011	-0.052	0.003
2023	0.614	0.487	0.925	-0.077	-0.003	-0.072	0.001
2024	0.672	0.549	0.903	-0.015	-0.004	-0.011	0.000

Table 26: Gemma 2 2B, Top 5 contributing *common temporal* nodes.

E Circuit tracing alternatives

E.1 EAP-IG Graphs

As a complementary study to understand the overlap of understanding between LLM component circuit tracing and transcoder circuit tracing we recreate the temporal head study (Park et al., 2025) for Gemma 2 2B, Qwen3-4b and LLaMA 3.2 1B, upgrading the EAP-IG package to accommodate group query attention. Our fork of Temporal Heads with upgraded EAP-IG algorithm can be found here: `placeholder`

Gemma 2 2B: Applying the methodology (Figure 13) we find A0H7, A2H3 and A7H4 as a common attention temporal heads. Examining the MLP components we note that the M14 node is common across time invariant and time sensitive questions. However, as EAP-IG circuit analysis is based on neural components, we are unable to understand the features and the 'why?' behind the importance of M14.

Qwen3-4b: There are no apparent temporal attention heads for Qwen3.

LLaMA 3.2 1B: There are no apparent temporal attention heads for LLaMA3.2.

E.2 SAE Circuit Tracing

We focussed on per layer transcoders as they produce richer representations and fewer error nodes leading to better interpretability and steering outcomes (Dunefsky et al., 2024). Examining a single question regarding the winner of the Basketball Bundesliga in 2022 and 2024 using Gemma 2 2B, SAEs (MLP, attention and residual stream saes) and Feature Circuits (Marks et al., 2025), we note a strong presence of residual stream and mlp features overlapping with our findings for higher temporal representations (Figure 14). However, we also note a lack of nodes at lower layers, such as 4-7, which have been identified in our exploration of transcoders as containing both *common temporal* and *common to the year* nodes. Whilst we have conducted preliminary investigations into the overlap of Feature Circuit with Circuit tracing our exploration is limited. Our fork of Feature Circuits can be found here: `placeholder`

F Optimised circuit tracing

Graphing differences for low confidence graphs Our initial conditions utilised high confidence outcomes, constructing circuits based on top-k logit outcomes. Whilst this provides a convenient way to interrogate for high confidence organic answers, for our purposes we want to understand the impact low confidence answers have on the influence scores, and the presentation of temporal nodes.

Low confidence graphs present similar year specific temporal features highlighting the strong contribution of temporal anchors even in low confidence settings. However, our selection of high confidence examples is justified if we examine the difference in temporal nodes between high and low confidence settings.

We note an increase in higher layer temporal representations (equal to or above layer 10) for our high confidence set, as well as an overall higher number of nodes. Surprisingly, average influence for all nodes per year across both sets are equivalent.

F.1 Simplified Temporal Representations

The indirect influence and node filtering from transcoder circuit tracing limit the number of nodes associated with the temporal component of our

	Overall Node Count		Higher Layer Node Count	
Year	High Conf.	Low Conf.	High Conf.	Low Conf.
2020	56	37	6	1
2021	52	50	3	3
2022	44	46	5	3
2023	81	46	6	2
2024	106	64	4	1
Common	28	30	N/A	N/A

Table 27: Node counts by year comparing overall and high layer representations.

test statements. We want to further understand the overlap between our findings and simplified temporal representations. To do so, we generate a simplified version of an explicit prompt (such as "In 2020"). Our investigation finds that the simplified time statements for Gemma 2 2B produce a strong overlap with the high confidence graphs, but do not entirely eclipse their representations. Focussing on the layer 4-14, there are obvious differences. Discrepancies were observed between the node structures of the simplified and original prompts. Notably, certain expected nodes are absent. Most significant are nodes such as L6F2246; while categorised under 'years and dates,' it also activates on semantic markers like 'coronavirus' and 'Edo period,' suggesting a broader or more entangled encoding of temporal context.

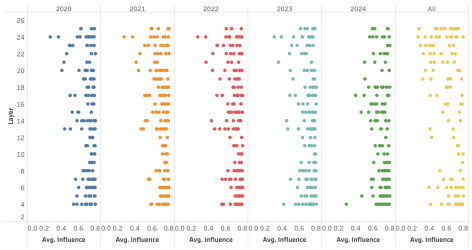


Figure 11: Simplified temporal representations for 2020–2024.

G Scaled Chrono-semantic Steering

We conducted a preliminary investigation into scaled steering methodologies for improved steering outcomes. Given the partially overlapping and distinct temporal features of Gemma 2 2B and Qwen3-4b we hypothesised that a scaled representation of these features may lead to improved outcomes.

To explore scaled representations we first base lined a set of high confidence questions; the president of France and the United States of America over the period 2015 to 2024. We select these questions because of their high accuracy, and this period as it covers a change in presidents and overlaps with the training cutoff period of both models. Our baseline in this case is the original steering method. For comparison, we investigated two alternative methods.

Average activation: The application of average activation of Chrono-semantic nodes for the year we wish to align to, at layers 12 and above.

Subtractive: The addition of Chrono-semantic nodes for the year we wish to align, with a selection of Chrono-semantic nodes for the year we did **not** wish to align to. Both using layers 12 and above.

The results are mixed. In some respects the *subtractive* intervention outperforms, producing marginally better results than the *Average activation* intervention, and in others the opposite is true.

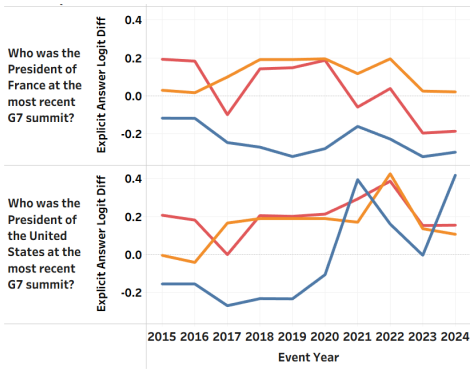


Figure 12: *Original steering* (blue), *average activation* (orange), *subtractive* (red). Demonstrates the improvement of the explicit answer logit probabilities with a scaled or selective approach, when compared to the original steering methodology

However, testing the *subtractive* intervention on the wider ChronosAlign dataset produces suboptimal outcomes (Figure 28 and 29). This suggests there is a dynamic effect between these Chrono-semantic nodes and the question at hand. A future direction of research could consider the relationship between questions, activations and answer probabilities; the development of a learned scaled representation of these nodes for specific outcomes.

	Metric	2020	2021	2022	2023	2024
F1	Steer F1	0.189	0.197	0.142	0.194	0.202
	Relative F1	0.071	0.085	0.020	0.046	0.081
	Explicit F1	0.351	0.370	0.416	0.320	0.148
	Invariant F1	0.083	0.076	0.056	0.089	0.118
Logit Diff	Explicit Logit Diff	0.031	0.064	0.016	0.045	0.023
	Relative Logit Diff	-0.028	-0.043	-0.029	-0.033	-0.034
	Invariant Logit Diff	-0.033	-0.050	-0.031	-0.038	-0.037
	AE Logit Diff	0.052	0.088	0.014	0.041	0.046
Probs. Change	Prior ive	0.073	0.079	0.076	0.074	0.045
	Prior rve	0.045	0.052	0.048	0.047	0.018
	Steered ive	0.009	-0.035	0.028	-0.009	-0.015
	Steered rve	-0.014	-0.055	0.003	-0.032	-0.039

Table 28: F1 scores, with logit probability difference for invariant and relative answers. Gemma 2 2B, Chronos-Align dataset using *Average activation* steering intervention with Chrono-semantic nodes at layer 12 and above.

	Metric	2020	2021	2022	2023	2024
F1	Steer F1	0.161	0.058	0.143	0.097	0.123
	Relative F1	0.071	0.085	0.020	0.046	0.081
	Explicit F1	0.351	0.370	0.416	0.320	0.148
	Invariant F1	0.083	0.076	0.056	0.089	0.118
Logit Diff	Explicit Logit Diff	0.036	0.015	0.042	0.012	-0.000
	Relative Logit Diff	-0.013	-0.019	-0.027	-0.018	-0.018
	Invariant Logit Diff	-0.018	-0.020	-0.029	-0.016	-0.016
	AE Logit Diff	0.029	0.015	0.035	0.006	0.005
Probs. Change	Prior ive	0.073	0.078	0.075	0.074	0.044
	Prior rve	0.045	0.051	0.048	0.046	0.017
	Steered ive	0.019	0.044	0.004	0.046	0.029
	Steered rve	-0.003	0.017	-0.021	0.016	-0.001

Table 29: F1 scores, with logit probability difference for invariant and relative answers. Gemma 2 2B, Chronos-Align dataset using *Subtractive* steering intervention with Chrono-semantic nodes, at layer 12 and above.

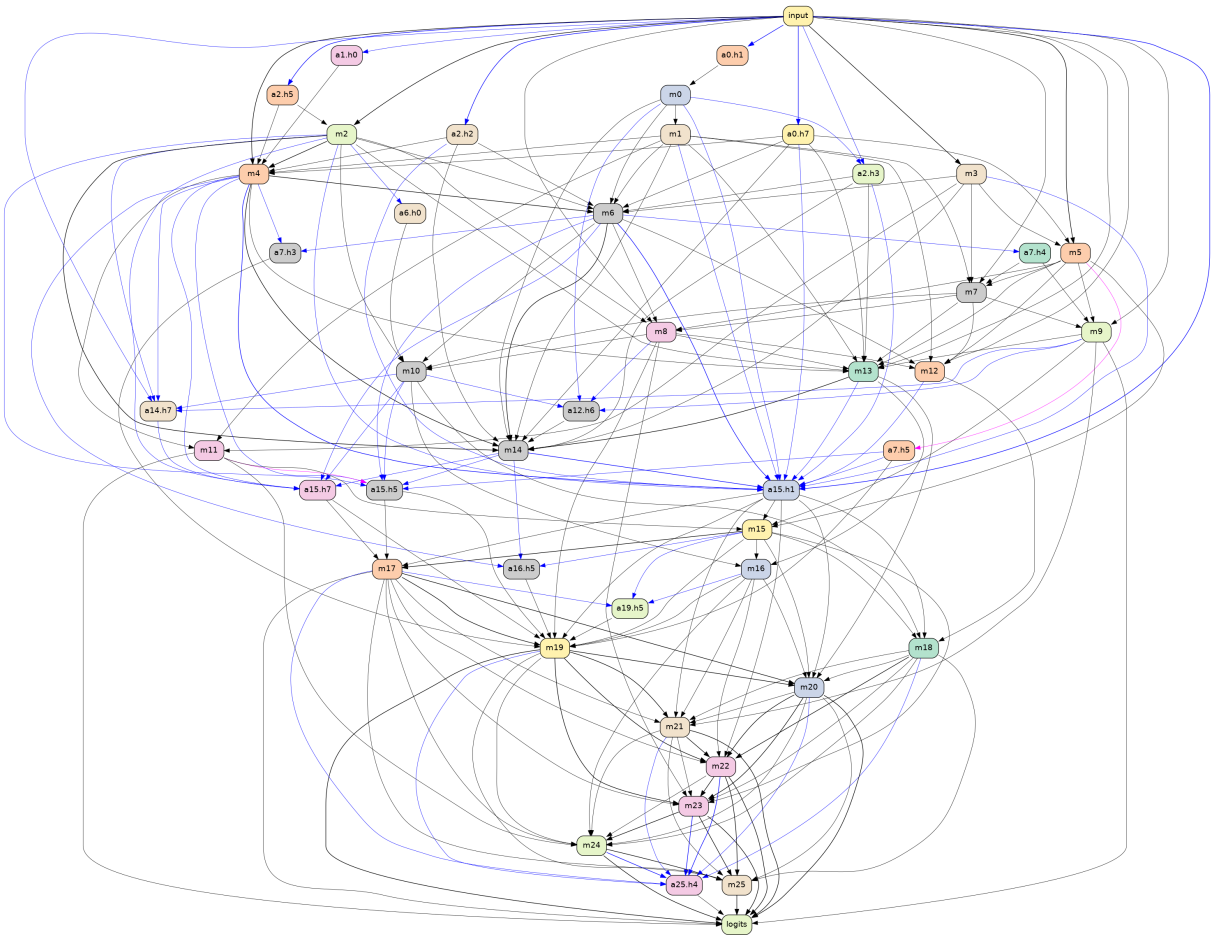


Figure 13: Gemma 2 sample EAP-IG circuit. Note the layer 7 head 4 attention, and M14 nodes



Figure 14: Gemma 2 2B SAE circuit tracing example showing residual stream and MLP feature interactions.