Automatic Feature Manifold Discovery in LLMs via Supervised Multi-Dimensional Scaling

Anonymous authors
Paper under double-blind review

Abstract

The linear representation hypothesis states that language models (LMs) encode concepts as directions in their latent space, forming organized, multidimensional manifolds. Prior efforts focus on discovering specific geometries for specific features, and thus lack generalization. We introduce Supervised Multi-Dimensional Scaling (SMDS), a model-agnostic method to automatically discover feature manifolds. We apply SMDS to temporal reasoning as a case study, finding that different features form various geometric structures such as circles, lines, and clusters. SMDS reveals many insights on these structures: they consistently reflect the properties of the concepts they represent; are stable across model families and sizes; actively support reasoning in models; and dynamically reshape in response to context changes. Together, our findings shed light on the functional role of feature manifolds, supporting a model of entity-based reasoning in which LMs encode and transform structured representations.[1]

1 Introduction

There is increasing evidence from recent work in mechanistic interpretability that language models (LMs) develop structured representations of entities in their latent space. Notably, Heinzerling & Inui (2024) find that numerical entities (e.g., **`Karl Popper was born in 1902`**) are represented in a monotonic, "pseudo-linear" fashion. Increasing or decreasing specific neuron activations can lead the model to output a higher or lower value. More recently, Engels et al. (2025) discover non-linear modes of structural entity representation, which form strikingly interpretable patterns. They show that days of the week (**`Sunday`**, **`Monday`**) and months (**`December`**, **`January`**), for example, form a circular structure. Concurrent work by Modell et al. (2025) provides formal definitions of these *feature manifolds* and explores how they arise in LMs.

Nevertheless, several fundamental questions remain unanswered: we do not know if and how LMs make use of these manifolds during reasoning, or how to reliably detect their presence (Engels et al., 2025; Modell et al., 2025). Answering these questions can help improve LMs and how we control them. This is particularly important in light of current LM limitations, such as poor temporal reasoning (Yuan et al., 2023), difficulty in alignment (Wang et al., 2023), bias (Gallegos et al., 2024), and vulnerability to distraction (Shi et al., 2023; Niu et al., 2025).

In this paper, we address these questions by introducing **Supervised Multi-Dimensional Scaling** (**SMDS**), a novel method to systematically discover feature manifolds. Unlike commonly used dimensionality reduction methods, which enforce a fixed structural assumption and cannot be directly compared, SMDS provides a unified way to specify arbitrary geometric assumptions and a quantitative metric to evaluate their fit. SMDS effectively turns manifold discovery into a model selection problem, and thus offers quantitative support for claims about the underlying structure of learned representations. Moreover, this method enables observing how a feature manifold evolves across different layers and reasoning steps.

We focus on temporal reasoning in the form of short-form QA tasks, such as identifying recency, ordering events and estimating durations, as we consider them an ideal test bed for manifold discovery. The reason

[1] Code and data will be publicly available online. *GitHub URL withdrawn for submission.*

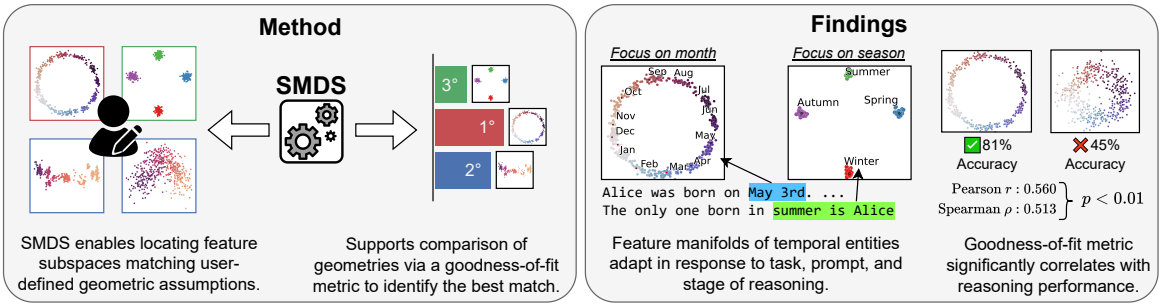

Figure 1: Our contributions. Supervised Multi-Dimensional Scaling is a novel dimensionality reduction technique to identify subspaces with a known geometry (left). Using it, we show evidence that temporal entities in LMs form various types of *feature manifolds*, which are task & prompt dependent and support reasoning (right).

is threefold: LMs display poor performance in such tasks (Yuan et al., 2023; Huang et al., 2023; Niu et al., 2024); initial evidence has found temporal feature manifolds to vary widely across tasks (Heinzerling & Inui, 2024; Engels et al., 2025); and finally, there is a gap in analyses targeting the *atomic structures* of temporal reasoning from a mechanistic standpoint.

The following are our main findings:

$\mathcal{F}_1$: Temporal entities form feature manifolds with intuitive structures, and this pattern is consistent across model architectures and sizes. We find that the manifolds associated with various temporal concepts (e.g., days of the year, hours, durations, and historical events) align with interpretable topologies such as circles, lines, and clusters, substantially extending Engels et al.'s (2025) findings. Our SMDS experiments cover more than sixty thousand recovered manifolds, and confirm that the identified feature structures are shared across model sizes and architectures.

$\mathcal{F}_2$: Feature manifolds are dynamically adjusted depending on the task. SMDS enables us to compare manifold structures across different token positions. We analyse prompts that share the same context but differ in their final completion cue, and find that LMs alter feature manifolds based on the cue and task in an intuitive way.

$\mathcal{F}_3$: Feature manifolds actively support reasoning. We find that LMs actively utilise feature manifolds to perform reasoning tasks, supported by two pieces of crucial evidence. First, perturbing manifold-aligned subspaces consistently impairs reasoning performance, while equivalent noise applied to random subspaces has a negligible effect. Second, we observe that manifold quality significantly correlates with downstream performance.

When combined with previous results on the binding problem (Feng & Steinhardt, 2023), our findings suggest an explanation for the mechanism by which LMs perform reasoning. We hypothesize an **entity-based reasoning pipeline** in LMs that:

1. Represents entity properties in coherent locations on a manifold within the residual stream;
2. Applies a transformation to this manifold, guided by the question or task context;
3. Selects an appropriate output based on the transformed representation.

Finally, we extend our analysis beyond mono-dimensional temporal features into two separate experiments (§5.4): the first is an entity-based reasoning task on geography that similarly uncovers manifold structures shared across models; the second studies a pair of temporal features to locate a multidimensional manifold. These experiments show our analysis can be extended beyond the temporal domain and to higher-dimensional

features. Overall, these results suggest that feature manifolds play an important role in how LMs represent and reason about entities. We view this work as a step toward better understanding the mechanisms behind reasoning in modern language models.

Contributions We first present a survey of previous feature manifold discovery methods, providing an overview of their limitations (§2). We then introduce the novel SMDS method in §3. Next, we present our results in §5, where we identify three major findings: (§5.1) manifold geometry for the same type of entity is shared across models; (§5.2) LMs adapt structures in context for different tasks; and (§5.3) LMs actively use feature manifolds for reasoning. Moreover, we show that our approach extends to other domains and to multidimensional manifolds (§5.4). Finally, we conclude the paper with a discussion (§6).

2 Feature Manifold Discovery

Existing methods for dimensionality reduction in manifold discovery often rely on fixed assumptions about the data distribution, without providing a principled way to compare results across different structural hypotheses. This gap motivates us to introduce our SMDS method in Section 3. In this section, we set up the problem with relevant background and survey existing feature manifold discovery methods.

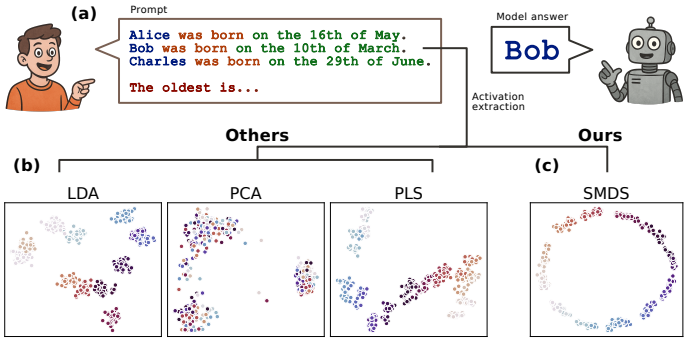

Figure 2: Feature Manifold Discovery and Limitations of Previous Methods.

Preliminaries We illustrate our method using a temporal reasoning task as a running example (Figure 2a). Performing temporal reasoning requires a model to understand both explicit mentions of temporal expressions (Jia et al., 2018b) and implicit knowledge of temporal calculus (Allen, 1981). Our analysis focuses on how LMs process temporal expressions, which are central to temporal reasoning and define precise, measurable quantities that can reveal underlying feature manifolds. Temporal reasoning also offers good diversity: different types of temporal expressions demand different reasoning skills (e.g., comparing frequencies, ordering events, or identifying recency) and models vary widely in how well they handle these tasks (Chu et al., 2024).

In particular, we seek to start from confirming Engels et al.'s (2025) finding that LMs tend to represent calendar dates in a circular topology, placing December near January in their latent space. Consider a prompt comprising several sentences following the template "`<name> was born on the <day> of <month>`." When asked "`The oldest is`," the task is answered correctly if the model uses contextual information to produce the correct answer `<name>`.

By prompting the LM with several such prompts varying the reference date, we elicit internal representations that collectively reside on the feature manifold of calendar dates. In this case, our quantity of interest is the birthday of the correct person (e.g., Bob's birthday: `10th of March` in Figure 2a), which we collectively represent as a set of labels y. We map these labels onto the $[0, 1]$ interval, where 0 corresponds to Jan 1st and 1 to Dec 31st. We then extract the hidden states corresponding to the last token of the date (e.g., the "`<month>`" token[2]), yielding a collection of hidden states $X \in \mathbb{R}^{n \times d}$, with n number of samples and d the hidden size of the LM. Next, we use dimensionality reduction to project the high-dimensional hidden states X onto an interpretable, low-dimensional space.

Existing Methods We identify three primary methods used in previous works: PCA, LDA, and PLS (Wold et al., 2001; Park et al., 2024a; El-Shangiti et al., 2025; Modell et al., 2025, *inter alia*).[3] From observing the visualisations in Figure 2b, we can see that each method has crucial limitations when trying to detect arbitrary geometries such as the circular one we seek. LDA finds interpretable clusters but has no notion of

[2]For readability, we omit space tokens in the examples. Tokenization is still performed as usual.
[3]We provide a review of relevant works in §A.

order; PCA fails to identify feature subspaces if they are not aligned with the directions of maximum variance; and PLS is limited to linear features unless a suitable transformation is applied to the data (AlquBoj et al., 2025). As a result, each method can only detect specific types of structure. Moreover, without quantitative metrics to assess the goodness-of-fit across different methods, it is unclear which of the manifolds best reflects the original representations.

3 Supervised Multi-Dimensional Scaling

To overcome these limitations, we propose a novel dimensionality reduction technique: Supervised Multi-Dimensional Scaling (SMDS). It extends classical Multi-Dimensional Scaling (MDS; Ghojogh et al., 2020) by incorporating supervision, under the assumption that labels can parametrise the underlying feature manifold formed by the model's hidden states. SMDS is flexible, as varying the assumption enables recovering multiple different structures, and provides a common basis to quantify their fit and identify a preferential one.

Formally, we assume that activations X forming the feature manifold can be located using labels y that represent a numerical property. SMDS first computes ideal pairwise distances $d(y_i, y_j)$ between $y_i, y_j \in y$ that encode the geometry of the desired manifold (e.g., circular, linear, or clustered). It then finds a linear projection $W \in \mathbb{R}^{m \times d}$ such that the Euclidean distances between projected points Wx_i and Wx_j best match $d(y_i, y_j)$, with $x_i, x_j \in X$. SMDS minimises the loss:

$$\mathcal{L} = \sum_{i<j} \left(\|W(x_i - x_j)\|^2 - d(y_i, y_j)^2 \right)^2. \tag{1}$$

$d(y_i, y_j)$ is task-dependent and implicitly defines the hypothesis structure. For example,

$$d(y_i, y_j) \coloneqq 2 \sin\left(\pi \min\left(\delta_{ij},\ 1 - \delta_{ij}\right)\right), \quad \delta_{ij} \coloneqq\ |y_i - y_j|, \tag{2}$$

these two formulas represent the chord distance between two points on a unit circle, thereby defining a *circular* structure. As shown in Figure 2c, SMDS finds a clear circular projection of calendar dates, consistent with Engels et al.'s (2025) findings.

We assess the quality of a recovered projection W trained on activations X by computing a variant of normalized stress (Amorim et al., 2014), adapted for a supervised task. In particular, we compute stress over a held-out set of points $\hat{X}, \hat{y}$ and corresponding ideal distances $\hat{d}_{ij} = d(\hat{y_i}, \hat{y_j})$:

$$S \coloneqq \sum_{i<j} \left[\|W\hat{x}_i - W\hat{x}_j\| - \hat{d}_{ij} \right]^2 / \sum_{i<j} \hat{d}_{ij}^2. \tag{3}$$

This metric measures how well distances in the recovered projection match distances of the hypothesis manifold. High-dimensional activations that originally form a certain structure can be easily projected onto a low-dimensional space matching that geometry, thus attaining a low stress. By comparing stress over several distance functions, one can identify the best-fitting manifold. For calendar dates, as we show later in §5, stress identifies a circular topology as the best fit among several different hypotheses.

Distance Functions We propose a set of distance functions for SMDS to detect a heterogeneous variety of manifolds. Seminal works have shown several instances of the idiosyncratic structure of feature manifolds. Notable examples include:

- Cyclical features form a ring shape in the latent space (Engels et al., 2025);
- Numbers are compressed according to a logarithmic progression (AlquBoj et al., 2025);
- Years of the 20th century form a U-shaped structure (Engels et al., 2025; Modell et al., 2025);
- Categorical features visually form clusters corresponding to the vertices of a polytope (Park et al., 2024a);
- Lastly, Gurnee & Tegmark (2023) have extracted multidimensional manifolds representing features such as latitude and longitude.

Table 2: Tasks and Corresponding Prompts. Variants `date_season`, `date_temperature`, and `time_of_day_phase` are omitted for brevity and are detailed in Appendix B. Colours represent templates: **blue** denotes names, **orange** denotes actions, **red** denotes the corresponding continuations, **green** denotes temporal expressions, and **black** denotes expressions that do not change throughout the dataset.

Dataset	Context	Continuation	Expression Range
date	Anna took a bus on the 16th of January.	The first person that took a bus was	01/01 - 31/12
duration	Neil is starting a workshop on the 11th of January lasting 1 day.	The person whose workshop ends first is	01/01 - 31/12 1 day - 4 years
notable	Emma was born on the day Pius X became Pope.	The oldest is	1900 - 2000
periodic	Kevin waters the plants every day.	The person who waters the plants more often is	daily - every 6 years
time_of_day	Lucy naps at 16:15.	It is now 19:37. The last person who napped is	00:00 - 23:59

Therefore, as listed in Table 1, we parametrise shapes such as circles, semicircles, lines, logarithmic lines and clusters so that the resulting manifold is interpretable. The manifolds we define are categorized based on their topology: linear, where concepts follow a continuous, monotonic progression; cyclical, where the progression is continuous but wraps around to the starting point, forming a loop; and categorical, where concepts occupy discrete, equidistant regions without inherent ordering.[4]

In the following sections, we use this collection of distance functions to identify feature manifolds for several tasks and at different stages of the reasoning process.

Table 1: Collection of distance functions used throughout our study. Colours denote manifold topology: **linear**, **cyclical** or **categorical**. $\delta_{ij} \coloneqq y_i - y_j$. $M \coloneqq \max(y)$.

Distance Function $d(y_i, y_j)$	Resulting Manifold
$\|\delta_{ij}\|$	linear
$\|\log y_i - \log y_j\|$	log_linear
$2\sin(\frac{\pi}{2}\|\delta_{ij}\|)$	semicircular
$2\sin(\frac{\pi}{2}\|\log y_i - \log y_j\|)$	log_semicircular
$2\sin(\pi \min(\|\delta_{ij}\|, 1 - \|\delta_{ij}\|))$	circular
$\min(\|\delta_{ij}\|, M + 1 - \|\delta_{ij}\|)$	discrete_circular
0 if $y_i = y_j$, 1 otherwise	cluster

4 Experimental Setup

Data & Prompt Setup Based on the TIMEX3 specification (Pustejovsky et al., 2010), we create five synthetic datasets and three variants, probing precise aspects of temporal understanding over a variety of numerical quantities (Table 2). All sentences across datasets have a similar format: they describe an action performed by three individuals, the action is associated with a temporal expression, and a continuation cue is attached to elicit temporal reasoning. The right answer is always one of the three names mentioned in the context. We randomise the names, actions, and temporal expressions to increase robustness but keep the same structure across all samples. Temporal expressions are sampled uniformly across a given range, but respecting some plausibility constraints (e.g. "**once per year**" is never associated with common actions such as "**takes a shower**"). We also make sure names are always tokenized as a single token for all models. The three variants `date_season`, `date_temperature`, and `time_of_day_phase` share the same context and range with their main counterparts, but ask a different question that requires a different type of reasoning (e.g. "**The only person born in spring is**"). Overall, our data exhibits greater variability than similar datasets used in previous literature. See Appendix B for an extended discussion on our temporal taxonomy, datasets and for the variability analysis.

LM Selection The bulk of our analysis is performed on three models from different families: Qwen2.5-3B-Instruct (Team et al., 2025), Llama-3.2-3B-Instruct (et al., 2024), gemma-2-2b-it (et al., 2025). We also study what impact instruction tuning has on these representations by comparing these models with their base versions. For the Llama family, we also study larger models to observe whether the manifolds we identify

[4]Strictly speaking, structures of this kind are not manifolds, as the space they form is not connected.

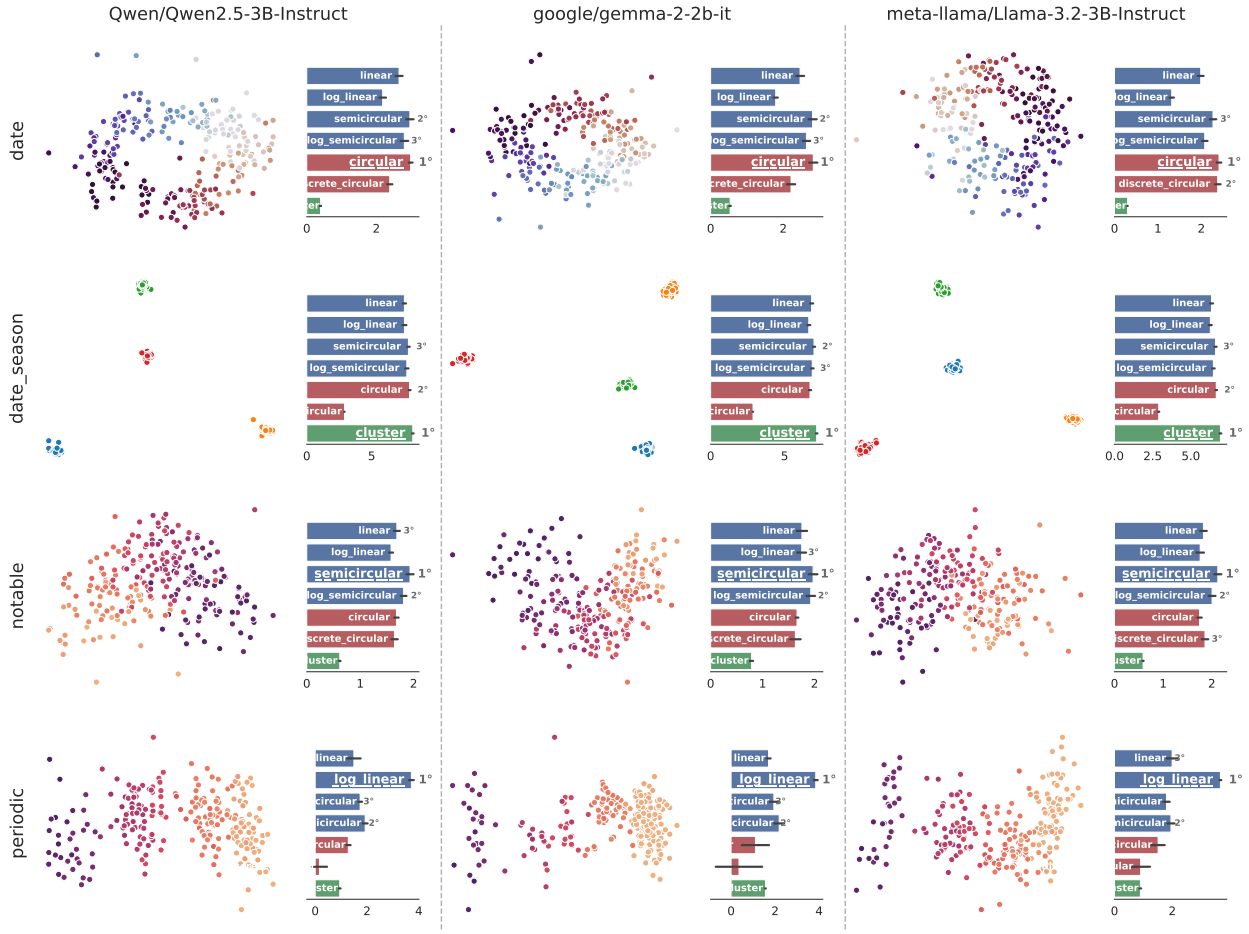

Figure 3: Feature manifolds retrieved from the LP site. We can observe that models represent features in a similar way, and the resulting manifolds are interpretable and match an intuitive progression (linear, circular or categorical) of the underlying features. The scatter plots on the **left** show the first two components of SMDS dimensionality reduction; the bar plots on the **right** depict scoring of different manifolds on the given activations. Scores displayed are computed as $-\log S$ to emphasise the difference between values; error bars are shown in black. Bar plot colour reflects manifold topology: linear; cyclical; categorical;

persist at scale: Llama-3.1-8B-Instruct and Llama-3.1-70B-Instruct. Due to computational constraints, we run these models using 4-bit quantization.

Automating Manifold Discovery We generalise our study by analysing activations across all layers and in different positions along the sentence. In particular, we examine three sites: the last token of the temporal expression (e.g., "`on the 16th of January`," TE for short); the last token in the prompt (e.g. "`The first person that took a bus was`," LP for short); and the token of the generated answer (one of the name of the context, A for short).

To automate the discovery process, we drop any assumption about which feature should be encoded by which manifold and instead run a grid search over all defined distance functions. We fit an instance of SMDS for each dataset and layer and then compare recovered manifolds using stress (Eq. 3). Throughout the study, we choose $m = 3$ as it is the minimum number of dimensions required to represent all our hypothesis manifolds (1D for most linear, 2D for some linear and cyclical manifolds, and 3D for clusters, which form a tetrahedron in 3D space). Higher values give similar results. Unless stated otherwise, all manifolds visualized in the study show the first two components identified by SMDS for the best-scoring layer, computed with a 50/50 train/test split. To increase the robustness of our manifold discovery, we perform 5-fold cross-validation,

training on four folds and evaluating S on the held-out fold in each rotation. We report the average S across all five folds as the overall score for the manifold.

5 Experiment Results and Analysis

We present our experiment results around the three major findings in this section.

5.1 ($\mathcal{F}_1$) Temporal Entities Share Intuitive Manifold Structures Across Models.

Figure 3, Table 3 show the best-scoring manifolds across models and tasks. We first observe that all manifolds identified this way are not only interpretable, but also match prior research (Engels et al., 2025; Park et al., 2024a; AlquBoj et al., 2025). Their topology always matches meaningful properties of the feature they explain: monotonic features are represented by linear topologies, cyclical features wrap around in loops, and categorical features map to cluster structures. Notably, the best manifold shape is consistent across all observed model families as well as in most of the non-instruction-tuned counterparts. Moreover, this pattern persists at scale, with all three observed sizes (3B, 8B, 70B) creating coherent shapes between them. This suggests there are preferential ways to encode the same knowledge, and all language models eventually converge to similar structures, providing further proof of hypotheses formulated in previous literature (Huh et al., 2024).

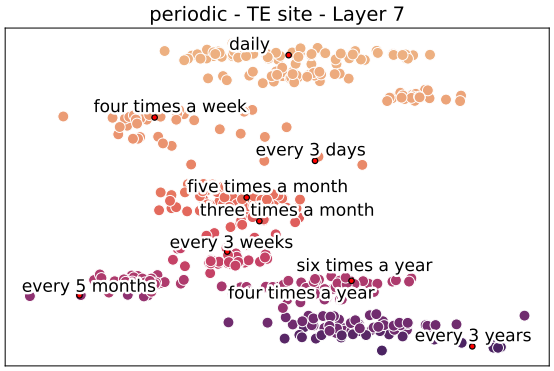

Figure 4: Llama-3.2-3B-Instruct on the `periodic` task. Events display logarithmic compression in their frequency: long intervals (e.g., months, years) are represented with the same granularity as shorter ones (e.g., days, weeks).

Previous work has shown that LMs encode numerical quantities in a logarithmically compressed way (AlquBoj et al., 2025). Our work extends this finding to temporal reasoning for the first time: in both the `duration` and `periodic` tasks, time intervals such as days to weeks, weeks to months, and months to years are preferentially

Table 3: Stress values for different models and tasks.

Model (instr-tuned)	**date**		**date_season**		**date_temperature**		**duration**	
	Best	$-\log S$	Best	$-\log S$	Best	$-\log S$	Best	$-\log S$
Llama-3.2-3B-Instruct	circ	2.40	cluster	6.93	cluster	6.39	log_lin	3.04
Qwen2.5-3B-Instruct	circ	2.98	cluster	8.16	log_smc	7.34	log_lin	2.76
gemma-2-2b-it	circ	2.79	cluster	7.16	log_smc	7.12	log_lin	3.33
Llama-3.1-70B-Instruct	circ	2.90	cluster	6.02	log_smc	6.19	log_lin	3.19
Llama-3.1-8B-Instruct	circ	2.41	cluster	6.50	cluster	5.92	log_lin	2.90
Model (base)								
Llama-3.2-3B	disc_circ	2.013	cluster	6.716	log_lin	6.683	euc	1.899
Qwen2.5-3B	circ	2.588	cluster	8.274	disc_circ	7.131	euc	1.815
gemma-2-2b	semicirc	3.166	cluster	7.277	disc_circ	6.423	euc	1.624

Model (instr-tuned)	**notable**		**periodic**		**time_of_day**		**time_of_day_phase**	
	Best	$-\log S$	Best	$-\log S$	Best	$-\log S$	Best	$-\log S$
Llama-3.2-3B-Instruct	semicirc	2.11	log_lin	3.66	circ	1.22	cluster	6.94
Qwen2.5-3B-Instruct	semicirc	1.92	log_lin	3.69	circ	1.14	cluster	8.55
gemma-2-2b-it	semicirc	1.95	log_lin	3.83	semicirc	1.22	cluster	7.38
Llama-3.1-70B-Instruct	semicirc	2.62	log_lin	3.85	circ	1.48	cluster	6.17
Llama-3.1-8B-Instruct	semicirc	2.20	log_lin	3.78	circ	1.26	cluster	6.58
Model (base)								
Llama-3.2-3B	cluster	0.93	log_lin	3.48	circ	1.24	cluster	6.92
Qwen2.5-3B	semicirc	1.82	log_lin	3.26	circ	1.28	cluster	8.61
gemma-2-2b	disc_circ	0.93	log_lin	3.63	circ	1.51	cluster	7.46

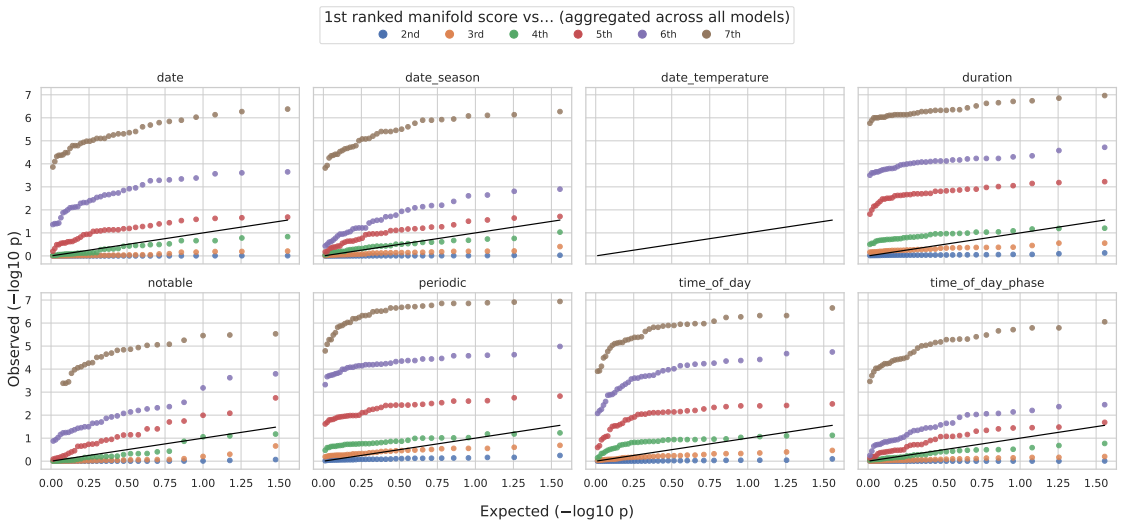

Figure 5: Q-Q plot of p-values obtained from a Nemenyi test comparing the stress scores of the first ranked manifold with all others for a given dataset. For all datasets except for `date_temperature`, the top 3-4 manifolds identified by SMDS score significantly better than the rest.

represented with roughly uniform spacing, indicating a logarithmic compression of temporal magnitude. This pattern (Figure 4), though not directly comparable, bears a superficial resemblance to the logarithmic compression described by the Weber–Fechner law (Dehaene, 2003). We note that the labels we use are themselves logarithmically spaced. A deeper study into temporal understanding is therefore needed to clarify whether the compression we observe is a genuine emergent behaviour or an artifact of our synthetic dataset.

Many tasks exhibit high scores for more than one topology. To clarify this ambiguity, we run a statistical significance study by performing a 10-fold cross-validation repeated 5 times across all datasets, models, and manifolds. First, we group observation across dataset, model, and rank of the manifold, and perform a Friedman test. Then, for all groups that achieve statistical significance ($p < 0.05$) we perform a post-hoc Nemenyi test to evaluate the significance of manifolds ranks on a given dataset. Figure 5 shows Q-Q plots of p-values grouped by dataset. Across most tasks, the manifolds ranked 1st performs comparably to manifolds ranked 2nd to 4th, and significantly better than the rest. For `duration`, `periodic`, and `time_of_day` this further reduces to manifolds 2nd to 3rd. The binary nature of `date_temperature` produces very homogeneous scores across all datasets and therefore we are not able to verify any significance. We discuss this further in Appendix D. This analysis both validates SMDS as a manifold discovery tool and confirms there is a small set of manifolds which consistently rank high on a given problem. A possible explanation is that, although preferential manifolds exist, models build multiple valid representations. This polymorphism is not an artefact of SMDS: control tasks using randomized labels display high stress, confirming SMDS is not just overfitting a hypothesis manifold. We provide a deeper discussion in Appendix D.5.

5.2 ($\mathcal{F}_2$) LMs Adapt Structures In-Context for Different Tasks.

This section describes two observed phenomena in which LMs reshape manifolds across tasks and depth.

Figure 6 shows how the LM adapts the TE site feature manifold to different structures at the LP site, depending on the question prompt. Tasks `date`, `date_season` and `date_temperature` all start from the same context but result in strikingly different final structures: in `date`, a circular structure is required to account for the looping nature of dates in a year, while in the other two tasks inputs are mapped to linearly separable clusters. This can be interpreted as the model internally performing regression or classification to solve the task.

When comparing the location in the sentence where the structure is located, models exhibit a form of information flow between entities, which can strengthen certain manifold structures, degrade others, or even drastically change their shape as observed earlier. Figure 7 shows how to detect this flow with stress. In initial layers, the TA site is highly structured. As layers progress, this structure disperses into later tokens, such as the LP token and the A token. This process is not perfect: duplicated manifolds on LP and A display

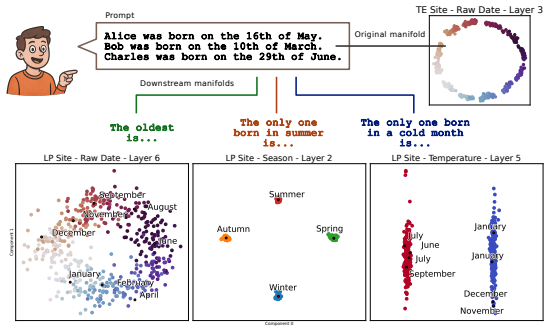

Figure 6: Feature manifolds of Llama-3.2-3B-Instruct on the `date` task and its variants. Best-scoring layers shown, as identified in Section 4. Different continuations produce drastically different topologies.

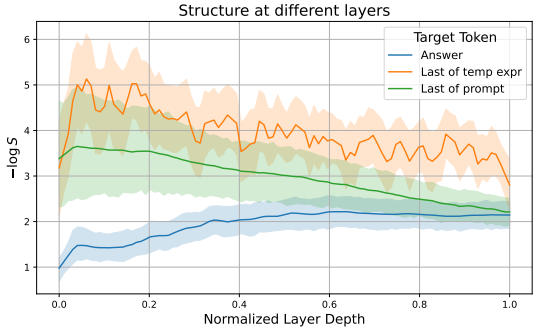

Figure 7: Manifold quality at different layers and positions in the sentence. Information transits from its injection point (orange) to the answer token (blue).

noticeably higher stress than the ones found at the TE site. A possibility is that later tokens in the same sentence accumulate more contextual information than early ones, thus resulting in noisier manifolds. Our results extend previous findings on the existence of a binding mechanism in LMs (Feng & Steinhardt, 2023; Dai et al., 2024): we show that not only vectors, but entire feature manifolds are preserved and propagated between entities.

5.3 ($\mathcal{F}_3$) LMs Actively Use Feature Manifolds for Reasoning.

Here we present two causally relevant lines of evidence that LMs actively use the structure of their representations to perform temporal reasoning.

Located subspaces are causally relevant to noise perturbation. To demonstrate that feature manifolds are utilized by LMs in their reasoning process, we perform causal intervention by adding noise to the manifold subspace and measuring downstream accuracy. We inject Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2 I_m)$ into the first layer at the TE site. Given a hidden state $x \in \mathbb{R}^d$, the perturbation is applied as $x' = x + W^{-1}\epsilon$, where ϵ is an m-dimensional noise vector projected back into the original space. Subspaces of dimension m are located via SMDS in the usual way, and overfitting is prevented by training and evaluating the SMDS on a 50/50 split. We select the top three task-model pairs achieving the best accuracy on the original task, as these will be the settings where a disruption will be more noticeable: `date`, `date_season` and `time_of_day_phase` on Llama-3.2-3B-Instruct.

Across all tasks, performance gracefully degrades as the noise scale is increased (Figure 8). Crucially, we observe degradation for m as low as 2, suggesting that temporal features are concentrated in very small yet highly informative regions of the activation space. We perform two other types of intervention in which we inject noise in the full latent space and in a random subspace, respectively. Affecting the full latent space achieves a much more destructive effect for low values of σ^2. On the other hand, disrupting a random subspace has no detectable effect on performance for subspaces of size < 100. The addition of noise also results in the disruption of structures located at subsequent tokens and layers (Figure 9). Interestingly, later layers are still able to form a vaguely organized shape, meaning information is partially being propagated or reconstructed. Our choice of perturbing the first layer is empirically motivated by the fact that intervention on later layers did not show as strong an effect. We hypothesize this is because information propagates quickly across tokens and layers, therefore the model is able to reconstruct a manifold from context tokens even if its source token has been disrupted. Overall, our experiments confirm that SMDS-located subspaces are critical for temporal understanding.

Manifold quality significantly correlates with model performance. We find a significant positive correlation between downstream accuracy and the ability of models to form well-organized manifolds, as quantified by $-\log S$ (Spearman's $\rho = 0.513$, $p = 0.0174$; Pearson's $r = 0.560$, $p = 0.0083$). Notably, this

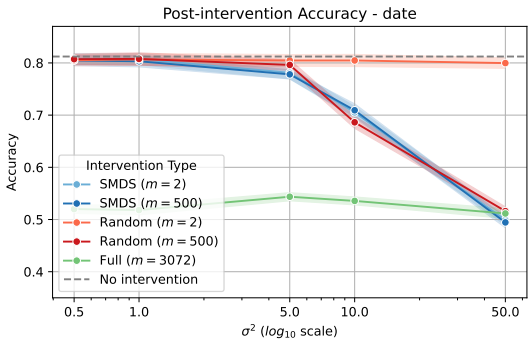

Figure 8: Downstream accuracy of Llama-3.2-3B-Instruct at increasing levels of noise. Error bars represent standard error. Intervening on a low-dimensional manifold subspace is just as effective at degrading model performance as perturbing much larger activation spaces.

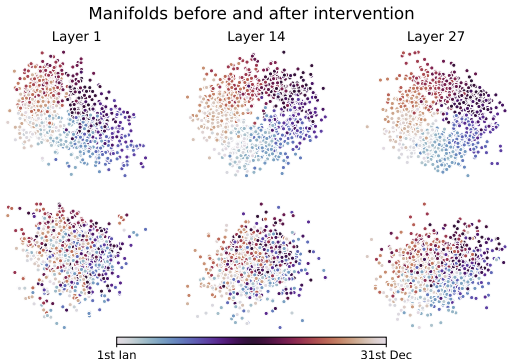

Figure 9: Llama-3.2-3B-Instruct on the `date` task. Latent space of the LP token before and after applying noise on the TE token (top and bottom respectively). Interventions on early tokens cause disruptions to the manifolds of later ones.

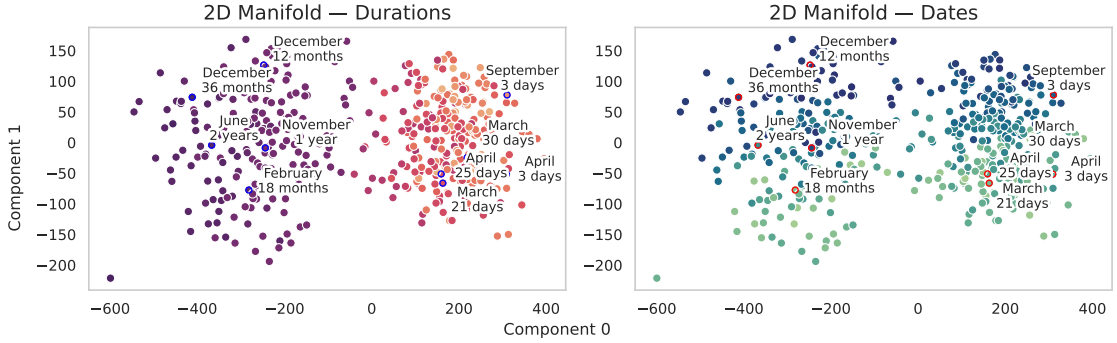

Figure 10: Multidimensional manifold constructed by Llama-3.2-3B-Instruct on the `duration` task. Component 0 is proportional to the duration, component 1 to the day of the year.

relationship emerges only for models that attain above-chance accuracy, specifically, Llama-3.2-3B-Instruct, Llama-3.1-8B-Instruct, and Llama-3.1-70B-Instruct. The results suggest that while feature manifolds tend to emerge naturally in LMs, a critical factor for strong performance lies in how effectively the model utilizes them during reasoning.

5.4 Generalising SMDS Across Domains and Feature Types

Defining manifolds through distance functions enables extending SMDS beyond the mono-dimensional case. This sections provides two such examples.

2D Manifold We analyse the `duration` task in more detail as each sentence contains two temporal expressions. We use both the duration and starting day to define a 2D label, and run SMDS with the **`linear`** hypothesis as it is the only distance function supporting multidimensional labels. Doing so reveals a 2D manifold that displays properties from both features. Table 4 presents Wilcoxon p-values obtained by comparing the model to a control task, showing they are significant across models. We hypothesize the creation of such manifolds happens during the information flow discussed earlier: features are retrieved from multiple locations and combined. The recovered manifold is shown in Figure 10.

Spatial Reasoning Domain To demonstrate the versatility of SMDS beyond temporal reasoning, we apply it to a manifold discovery task grounded in geographic knowledge. We construct a dataset of prompts referencing various cities around the world and use their latitude and longitude to compute pairwise distances and reconstruct a manifold. While Gurnee & Tegmark (2023) demonstrates that geographic location is decodable from LMs' hidden states, their analysis is limited to a planar projection. We extend this by evaluating spherical, cylindrical, and geodesic-based geometries, and find that a spherical manifold best captures the structure of the representations. This again highlights how feature manifolds align with the true geometry of the underlying domain. Further details are provided in Appendix D.4.

Table 4: Stress values for the `duration` task at the LT site evaluated with the **`linear`** manifold. Standard error shown in grey. Score differences with a control task with randomized labels are all statistically significant.

Model	Best Layer	$-\log S$	Control $-\log S$	p
Llama-3.2-3B-Instruct	20	$2.173_{\pm 0.114}$	$0.469_{\pm 0.020}$	0.031
Qwen2.5-3B-Instruct	2	$2.585_{\pm 0.047}$	$0.506_{\pm 0.024}$	0.031
gemma-2-2b-it	2	$2.595_{\pm 0.065}$	$0.519_{\pm 0.014}$	0.031

6 Discussion & Conclusion

Table 5: Stress values for the `cities` task at the RC site. The highest-scoring manifold is always a spherical one.

Model	Acc	Manifold $-\log S$ cylinder	flat	geodesic	sphere
Llama-3.2-3B-it	0.549	2.071	1.931	2.118	**2.285**
Qwen2.5-3B-it	0.510	1.906	1.768	1.975	**2.135**
gemma-2-2b-it	0.493	2.070	1.947	2.073	**2.248**

Our study establishes a connection between the geometry of representation manifolds and the causal language modelling process, demonstrating that a structured organization of knowledge is not only present but beneficial for model reasoning. By analysing the persistence of these structures across tokens—particularly from the injection point to the answer—we provide compelling evidence that feature binding operates through continuous, task-relevant manifolds in the latent space. The persistence of manifolds across tokens suggests that language models transfer not just vectors, but structured representations, reinforcing the presence of a binding mechanism and extending prior evidence to more diverse tasks (Dai et al., 2024).

Although our experiments centre on temporal reasoning, the proposed method extends to any task involving structured features on which a distance function can be defined, as we demonstrate in §D.4. Starting from hypothesis manifolds inspired by prior work, we obtain consistent, interpretable results, effectively reframing manifold discovery as a model selection problem. A compelling direction for future research is understanding how individual features combine into multidimensional manifolds. While we present initial evidence of composition, more expressive manifold hypotheses could offer deeper insights. SMDS lays the foundation for such investigations.

Our stress metric often yields tightly clustered scores. This suggests three possible directions for further investigation: the development of more discriminative metrics, the use of larger and more varied datasets, or a reassessment of the assumption that a single preferential manifold exists. An intriguing hypothesis is that models instead adopt multiple, equally valid representational geometries and dynamically select them based on task context.

In the scope of model reasoning, manifold discovery can serve as a basis for several lines of future work. For instance, combining SMDS with circuit discovery Conmy et al. (2023) could help identify which operations LMs use to transform information throughout reasoning. Another promising direction is model steering Park et al. (2024b), where knowledge of feature manifolds could inform methods that leverage these structures directly. Finally, systematically studying the role of noise in feature manifolds across layers, and whether mitigating it improves reasoning, offers another rich line of inquiry.

In sum, *shape happens*. Our work lays the foundational ground for interpreting and comparing representations in LMs through geometric structures. This invites further exploration into how manifold shapes are formed, combined, functionally employed in downstream reasoning, and how knowing about them could improve existing models.

Limitations

Our use of language models trained on predominantly English corpora introduces an inherent bias toward the cultural norms of the Anglosphere. This is reflected in several design choices: the reliance on the Gregorian calendar for date expressions; the selection of names that are tokenized as single units, which tends to privilege Anglo-American names; and assumptions about seasonal properties (e.g., associating December with cold weather), which implicitly expects the location to be a country in the northern hemisphere, with a temperate or continental climate. The high accuracy and well-formed manifolds observed in these settings can therefore be seen as indicators of such biases. SMDS could find use as a diagnostic tool, uncovering how underlying representations reflect these biases.

In our work, we omit fuzzy expressions for which it is not possible to define precise temporal pointers (e.g., "in the morning," "later," and "next week") and therefore an exact location on a feature manifold. As Kenneweg et al. (2025) show, fuzziness in a temporal expression is a key factor in performance degradation. Future works could better characterize the interplay between fuzziness in temporal expressions and the quality of feature manifolds.

References

James F. Allen. An interval-based representation of temporal knowledge. In *In Proceedings 7th IJCAI*, pp. 221–226, 1981.

H. V. AlquBoj, Hilal AlQuabeh, Velibor Bojkovic, Tatsuya Hiraoka, Ahmed Oumar El-Shangiti, Munachiso Nwadike, and Kentaro Inui. Number Representations in LLMs: A Computational Parallel to Human Perception, February 2025.

Elisa Amorim, Emilio Vital Brazil, Luis Gustavo Nonato, Faramarz Samavati, and Mario Costa Sousa. Multidimensional Projection with Radial Basis Function and Control Points Selection. In *2014 IEEE Pacific Visualization Symposium*, pp. 209–216, March 2014. doi: 10.1109/PacificVis.2014.59.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the Cross-lingual Transferability of Monolingual Representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.421.

Yonatan Belinkov. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*, 48(1):207–219, March 2022. doi: 10.1162/coli_a_00422.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL `https://transformer-circuits.pub/2023/monosemantic-features/index.html`.

Tyler Chang, Zhuowen Tu, and Benjamin Bergen. The Geometry of Multilingual Language Model Representations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 119–136, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.9.

David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, Satvik Golechha, and Joseph Bloom. A is for Absorption: Studying Feature Splitting and Absorption in Sparse Autoencoders, June 2025.

Wenhu Chen, Xinyi Wang, William Yang Wang, and William Yang Wang. A Dataset for Answering Time-Sensitive Questions. In J. Vanschoren and S. Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Haotian Wang, Ming Liu, and Bing Qin. TimeBench: A Comprehensive Evaluation of Temporal Reasoning Abilities in Large Language Models. In

Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1204–1228, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.66.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tai, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. *J. Mach. Learn. Res.*, 25(1):70:3381–70:3433, January 2024. ISSN 1532-4435.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Thirty-Seventh Conference on Neural Information Processing Systems*, 2023.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. Emerging Cross-lingual Structure in Pretrained Language Models. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6022–6034, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.536.

Qin Dai, Benjamin Heinzerling, and Kentaro Inui. Representational Analysis of Binding in Language Models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17468–17493, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.967.

Stanislas Dehaene. The neural basis of the Weber–Fechner law: A logarithmic mental number line. *Trends in Cognitive Sciences*, 7(4):145–147, April 2003. ISSN 1364-6613. doi: 10.1016/S1364-6613(03)00055-X.

Ahmed Oumar El-Shangiti, Tatsuya Hiraoka, Hilal AlQuabeh, Benjamin Heinzerling, and Kentaro Inui. The Geometry of Numerical Reasoning: Language Models Compare Numeric Properties in Linear Subspaces. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 550–561, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-190-2. doi: 10.18653/v1/2025.naacl-short.47.

Joshua Engels, Eric J Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear. In *The Thirteenth International Conference on Learning Representations*, 2025.

Gemma et al. Gemma 3 Technical Report, March 2025.

Grattafiori et al. The Llama 3 Herd of Models, November 2024.

Eoin Farrell, Yeu-Tong Lau, and Arthur Conmy. Applying Sparse Autoencoders to Unlearn Knowledge in Language Models. In *Neurips Safe Generative AI Workshop 2024*, October 2024.

Bahare Fatemi, Mehran Kazemi, Anton Tsitsulin, Karishma Malkan, Jinyeong Yim, John Palowitch, Sungyong Seo, Jonathan Halcrow, and Bryan Perozzi. Test of Time: A Benchmark for Evaluating LLMs on Temporal Reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025.

Jiahai Feng and Jacob Steinhardt. How do Language Models Bind Entities in Context? In *The Twelfth International Conference on Learning Representations*, October 2023.

Lisa Ferro, Laurie Gerber, Inderjeet Mani, Beth M. Sundheim, and George Wilson. 2003 Standard for the Annotation of Temporal Expressions. 2004.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics*, 50(3):1097–1179, September 2024. doi: 10.1162/coli_a_00524.

Benyamin Ghojogh, Ali Ghodsi, Fakhri Karray, and Mark Crowley. Multidimensional Scaling, Sammon Mapping, and Isomap: Tutorial and Survey, September 2020.

Sreyan Ghosh, Chandra Kiran Reddy Evuru, Sonal Kumar, Ramaneswaran S, Deepali Aneja, Zeyu Jin, Ramani Duraiswami, and Dinesh Manocha. A closer look at the limitations of instruction tuning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *ICML'24*, pp. 15559–15589, Vienna, Austria, July 2024. JMLR.org.

Wes Gurnee and Max Tegmark. Language Models Represent Space and Time. In *The Twelfth International Conference on Learning Representations*, October 2023.

John Healy and Leland McInnes. Uniform manifold approximation and projection. *Nature Reviews Methods Primers*, 4(1):82, November 2024. ISSN 2662-8449. doi: 10.1038/s43586-024-00363-x.

Benjamin Heinzerling and Kentaro Inui. Monotonic Representation of Numeric Attributes in Language Models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 175–195, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-short.18.

John Hewitt and Percy Liang. Designing and Interpreting Probes with Control Tasks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1275.

Quzhe Huang, Yutong Hu, Shengqi Zhu, Yansong Feng, Chang Liu, and Dongyan Zhao. More than Classification: A Unified Framework for Event Temporal Relation Extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9631–9646, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.536.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse Autoencoders Find Highly Interpretable Features in Language Models. In *The Twelfth International Conference on Learning Representations*, October 2023.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The Platonic Representation Hypothesis. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 20617–20642. PMLR, July 2024.

Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, and Gerhard Weikum. TEQUILA: Temporal Question Answering over Knowledge Bases. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, pp. 1807–1810, New York, NY, USA, 2018a. Association for Computing Machinery. ISBN 978-1-4503-6014-2. doi: 10.1145/3269206.3269247.

Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, and Gerhard Weikum. TempQuestions: A Benchmark for Temporal Question Answering. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, pp. 1057–1062, Republic and Canton of Geneva, CHE, 2018b. International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-5640-4. doi: 10.1145/3184558.3191536.

Zhen Jia, Soumajit Pramanik, Rishiraj Saha Roy, and Gerhard Weikum. Complex Temporal Question Answering on Knowledge Graphs. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, pp. 792–802, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 978-1-4503-8446-9. doi: 10.1145/3459637.3482416.

Subhash Kantamneni, Joshua Engels, Senthooran Rajamanoharan, Max Tegmark, and Neel Nanda. Are Sparse Autoencoders Useful? A Case Study in Sparse Probing. June 2025. URL https://openreview.net/forum?id=rNfzT8YkgO&noteId=hA0Ptt2IOa.

Svenja Kenneweg, Jörg Deigmöller, Philipp Cimiano, and Julian Eggert. TRAVELER: A Benchmark for Evaluating Temporal Reasoning across Vague, Implicit and Explicit References. *ArXiv*, abs/2505.01325, 2025.

Amit Arnold Levy and Mor Geva. Language Models Encode Numbers Using Digit Representations in Base 10. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 385–395, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-190-2. doi: 10.18653/v1/2025.naacl-short.33.

Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task. In *The Eleventh International Conference on Learning Representations*, September 2022.

Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. The Unlocking Spell on Base LLMs: Rethinking Alignment via In-Context Learning. In *The Twelfth International Conference on Learning Representations*, October 2023.

Raja Marjieh, Veniamin Veselovsky, Thomas L. Griffiths, and Ilia Sucholutsky. What is a Number, That a Large Language Model May Know It?, February 2025.

Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse Feature Circuits: Discovering and Editing Interpretable Causal Graphs in Language Models. In *The Thirteenth International Conference on Learning Representations*, October 2024.

Gouki Minegishi, Hiroki Furuta, Yusuke Iwasawa, and Yutaka Matsuo. Rethinking Evaluation of Sparse Autoencoders through the Representation of Polysemous Words. In *The Thirteenth International Conference on Learning Representations*, October 2024.

Alexander Modell, Patrick Rubin-Delanchy, and Nick Whiteley. The Origins of Representation Manifolds in Large Language Models, May 2025.

Jingcheng Niu, Saifei Liao, Victoria Ng, Simon De Montigny, and Gerald Penn. ConTempo: A Unified Temporally Contrastive Framework for Temporal Relation Extraction. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 1521–1533, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.89.

Jingcheng Niu, Xingdi Yuan, Tong Wang, Hamidreza Saghir, and Amir H. Abdi. Llama See, Llama Do: A Mechanistic Perspective on Contextual Entrainment and Distraction in LLMs, June 2025.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The Geometry of Categorical and Hierarchical Concepts in Large Language Models. In *ICML 2024 Workshop on Mechanistic Interpretability*, June 2024a.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *ICML'24*, pp. 39643–39666, Vienna, Austria, July 2024b. JMLR.org.

Qiwei Peng and Anders Søgaard. Concept Space Alignment in Multilingual LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5511–5526, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.315.

James Pustejovsky, Kiyong Lee, Harry Bunt, and Laurent Romary. ISO-TimeML: An International Standard for Semantic Annotation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010. European Language Resources Association (ELRA).

Roser Saurí, Jessica Moszkowicz, Bob Knippen, Rob Gaizauskas, Andrea Setzer, and James Pustejovsky. TimeML Annotation Guidelines Version 1.2.1. January 2006.

Andrea Setzer. *Temporal Information in Newswire Articles: An Annotation Scheme and Corpus Study.* PhD dissertation, University of Sheffield, 2001.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, Stella Biderman, Adria Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, Eric J. Michaud, Stephen Casper, Max Tegmark, William Saunders, David Bau, Eric Todd, Atticus Geiger, Mor Geva, Jesse Hoogland, Daniel Murfet, and Tom McGrath. Open Problems in Mechanistic Interpretability, January 2025.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. Large Language Models Can Be Easily Distracted by Irrelevant Context. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 31210–31227. PMLR, July 2023.

Juan Luis Suárez-Díaz, Salvador García, and Francisco Herrera. A Tutorial on Distance Metric Learning: Mathematical Foundations, Algorithms, Experimental Analysis, Prospects and Challenges (with Appendices on Mathematical Background and Detailed Algorithms Explanation), August 2020.

Varshini Subhash, Anna Bialas, Weiwei Pan, and Finale Doshi-Velez. Why do universal adversarial attacks work on large language models?: Geometry might be the answer. In *The Second Workshop on New Frontiers in Adversarial Machine Learning*, August 2023.

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. Towards Benchmarking and Improving the Temporal Reasoning Capability of Large Language Models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14820–14835, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.828.

Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 Technical Report, January 2025.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024.

Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. ISSN 1533-7928.

Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. Aligning Large Language Models with Human: A Survey, July 2023.

Yuqing Wang and Yun Zhao. TRAM: Benchmarking Temporal Reasoning for Large Language Models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6389–6415, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.382.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned Language Models are Zero-Shot Learners. In *International Conference on Learning Representations*, October 2021.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of*

the 36th International Conference on Neural Information Processing Systems, NIPS '22, pp. 24824–24837, Red Hook, NY, USA, November 2022. Curran Associates Inc. ISBN 978-1-7138-7108-8.

Svante Wold, Michael Sjöström, and Lennart Eriksson. PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2):109–130, October 2001. ISSN 0169-7439. doi: 10.1016/S0169-7439(01)00155-1.

Xuansheng Wu, Wenlin Yao, Jianshu Chen, Xiaoman Pan, Xiaoyang Wang, Ninghao Liu, and Dong Yu. From Language Modeling to Instruction Following: Understanding the Behavior Shift in LLMs after Instruction Tuning. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2341–2369, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.130.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. AxBench: Steering LLMs? Even Simple Baselines Outperform Sparse Autoencoders. In *Forty-Second International Conference on Machine Learning*, June 2025.

Chenhan Yuan, Qianqian Xie, and Sophia Ananiadou. Zero-shot Temporal Relation Extraction with ChatGPT, April 2023.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*, September 2019.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. "Going on a vacation" takes longer than "Going for a walk": A Study of Temporal Commonsense Understanding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/d19-1332.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. LIMA: Less Is More for Alignment. In *Thirty-Seventh Conference on Neural Information Processing Systems*, November 2023.

A Related Works

Linear representations & feature manifolds The linear representation hypothesis proposes that language models encode interpretable features as directions in their latent space, with concepts expressed as sparse linear combinations of these directions (Park et al., 2024b; Modell et al., 2025). Recent work extends this view, revealing that related features tend to organize into structured manifolds. For example, ring-like structures (Engels et al., 2025), logarithmic progressions (AlquBoj et al., 2025), U-shaped curves (Engels et al., 2025; Modell et al., 2025), clusters organized around the vertices of geometric polytopes (Park et al., 2024a), and higher-dimensional surfaces (Gurnee & Tegmark, 2023). Other studies on multilingual LMs have also investigated structures in the latent space and consistently found shared representations across languages (Peng & Søgaard, 2024; Artetxe et al., 2020; Chang et al., 2022; Conneau et al., 2020, *inter alia*).

Existing dimension reduction methods Several linear dimensionality reduction techniques have been applied to recover structure from language model representations: Principal Component Analysis (PCA) identifies directions of maximal variance in the embedding space (Gurnee & Tegmark, 2023; Modell et al., 2025); Linear Discriminant Analysis (LDA) finds directions that best separate labeled categories (Park et al., 2024a); Partial Least Squares Regression (PLS) identifies components that most strongly covary with target labels (Wold et al., 2001; El-Shangiti et al., 2025; Heinzerling & Inui, 2024); and Multi-Dimensional Scaling (MDS) seeks low-dimensional embeddings that preserve pairwise distances from the original space (Marjieh et al., 2025). In addition to these linear methods, some non-linear techniques such as t-SNE and UMAP have also been applied (van der Maaten & Hinton, 2008; Healy & McInnes, 2024; Subhash et al., 2023).

Another prominent technique used in interpretability works is the Sparse Auto Encoder (SAE), a neural network building a mapping from the dense activation space of a LM to a high-dimensional, sparse, latent space such that single neurons of a SAE represent atomic concepts (Bricken et al., 2023; Huben et al., 2023). SAEs have been successful at recovering vast collections of monosemantic, interpretable features at scale (Templeton et al., 2024), but have also found usage in unlearning (Farrell et al., 2024), detecting internal causal graphs (Marks et al., 2024) and identifying circuits (Minegishi et al., 2024). While SAEs have shown promise for LLM interpretability, they face substantial critiques and limitations that challenge their effectiveness and reliability. Representations identified by SAEs may fall victim of "feature absorption," complicating the disentanglement of atomic features (Chanin et al., 2025). In model steering, simple baselines have been observed outperforming SAEs (Wu et al., 2025; Kantamneni et al., 2025). Lastly, SAEs are expensive to construct as they require extremely large dimension of their latent space and necessitate in some cases billions of tokens for training. Their construction also makes them model-specific, preventing transferability (Sharkey et al., 2025).

In this work, we primarily compare SMDS with other linear dimensionality reduction techniques. While non-linear methods and SAEs may offer valuable insights, we do not focus on them here due to their high computational demands. Our goal is to enable a scalable and systematic exploration of feature manifolds across models and tasks. This requires lightweight methods that can be efficiently applied in closed form, making linear approaches better suited to the scope of our investigation.

Temporal reasoning Temporal reasoning refers to the ability to interpret and manipulate expressions that describe temporal information — such as dates (e.g., "May 1, 2010"), times ("9 pm"), or temporal relations ("before," "in the morning")—in order to determine when events occur or how they relate temporally (Jia et al., 2018a). Such reasoning tasks often require composing multiple temporal expressions to answer nuanced, time-sensitive questions.

Several datasets exist that seek to benchmark LMs across different facets of temporal reasoning. Some evaluate factual recall over time (Jia et al., 2018b; Chen et al., 2021; Jia et al., 2021), others focus on temporal understanding of real-world scenarios (Zhou et al., 2019; Fatemi et al., 2025), and yet others probe the temporal arithmetic capabilities of LMs (Tan et al., 2023). Lastly, some works have aggregated existing benchmarks in order to evaluate broader capabilities such as symbolic, commonsense, and event reasoning (Wang & Zhao, 2024; Chu et al., 2024).

Existing benchmarks primarily assess overall performance on complex tasks involving multiple temporal expressions and reasoning types. Simpler tasks focusing on specific types of temporal expressions, despite being foundational to temporal understanding, remain underexplored. To enable a mechanistic investigation of how language models process temporal information, homogeneous datasets that isolate specific facets of temporal expressions are required.

B Temporal Taxonomy & Datasets

This section describes the synthetic datasets we have generated to probe atomic aspects of temporal understanding.

Taxonomy Various annotation schemes have been developed to characterise temporal expressions such as TIMEX3 (Pustejovsky et al., 2010), TIMEX2 (Ferro et al., 2004), TIMEX (Setzer, 2001) and TimeML (Saurí et al., 2006), as well as several variants. We take inspiration from TIMEX1-3 to construct several synthetic datasets. Each one covers a specific family of temporal expressions (Table 2):

- `date`: Refers to a specific calendar date. To explore periodic reasoning, we omit the year;
- `time_of_day`: Specifies a precise moment in the day;
- `duration`: Defines a duration and its starting point;
- `periodic`: Refers to events that recur with a given frequency;

Table 6: Additional examples for each task.

Dataset	# Samples	Examples
cities	2000	Luke lives in Boston. William lives in Toronto. Michael lives in Cancún. The person who lives closest to Luke is
		Mark lives in Leuven. Jack lives in Heidelberg. Dallas lives in Messina. The person who lives closest to Mark is
date	1992	Brandon donated clothes on the 29th of September. Bob donated clothes on the 31st of August. Jerry donated clothes on the 27th of September. The first person that donated clothes was
		Matt visited a new city on the 22nd of February. Josh visited a new city on the 14th of February. Frank visited a new city on the 1st of March. The first person that visited a new city was
date_season	2000	Emily mowed the lawn on the 8th of December. Blake mowed the lawn on the 30th of April. Walker mowed the lawn on the 27th of June. The only person that mowed the lawn in fall is
		Rose painted a mural on the 16th of June. Robert painted a mural on the 13th of July. Martin painted a mural on the 27th of July. The only person that painted a mural in spring is
date_temperature	2000	Richard left for vacation on the 25th of June. Neil left for vacation on the 22nd of December. April left for vacation on the 22nd of August. The only person that left for vacation in a cold month is
		Jason returned from vacation on the 12th of February. Connor returned from vacation on the 21st of March. Rachel returned from vacation on the 19th of October. The only person that returned from vacation in a warm month is
duration	3000	Maria is starting their internship on the 15th of December and is set to run for 25 days. George is starting their internship on the 13th of December and is set to run for 14 days. Laura is starting their internship on the 3rd of December and is set to run for 1 week. The person whose internship ends first is
		Hunter runs a festival booth on the 27th of December staying open for 10 days. George runs a festival booth on the 12th of November staying open for 9 days. Connor runs a festival booth on the 20th of December staying open for 9 days. The person whose festival booth ends first is
notable	2000	Robert was born on the day the MV Doña Paz sank. Maria was born on the day the independent State of Palestine was proclaimed. Andrew was born on the day the Dayton Accords were signed. The oldest is
		Neil was born on the day Herbert Hoover was inaugurated as President. Leon was born on the day James Joyce published Ulysses. Alice was born on the day Mandatory Palestine was established. The oldest is
time_of_day	3000	Steve watches a movie at 23:15. April watches a movie at 11:45. Charlie watches a movie at 7:15. It is now 2:58. The last person who watched a movie is
		Charlie watches TV at 4:45. Richard watches TV at 12:15. Steve watches TV at 4:30. It is now 14:42. The last person who watched TV is
time_of_day_phase	2000	Leon goes for a walk at 6:15. Brandon goes for a walk at 18:30. Matt goes for a walk at 5:00. The only person that goes for a walk in the morning is
		John writes in a journal at 4:45. Matt writes in a journal at 5:00. Luke writes in a journal at 20:45. The only person that writes in a journal in the evening is

- notable: Contains an indirect but precise reference to an event taking place in a given moment in time;

The taxonomy has been defined in such a way that temporal expressions have a unique, precisely-defined associated numerical quantity. We have chosen to omit fuzzy expressions for which it is not possible to define precise temporal pointers (e.g. “in the morning”, “later”, “next week”) and therefore an exact position in a feature manifold.

Dataset Creation We build each sentence in the dataset by combining three contextual sentences and a termination that elicits reasoning. Each sentence contains a name, action and temporal expressions which are all uniformly sampled from a given set. Names and actions have been generated via ChatGPT and checked manually to be consistently formatted and the resulting sentences grammatically correct. Names have been chosen so that they are not broken up into separate tokens. Temporal expressions of the notable

Table 7: Possible temporal expressions for each task.

Dataset	**Temporal expression set**
`cities`	Uniformly sampled based on location from the World Cities Database, considering only prominent cities or cities with $>$ 10.000 inhabitants for US and Canada.
`date`, `date_season`, `date_temperature`	Uniformly sampled from all 365 days of a non-leap year.
`duration`	Dates sampled in the same way as `date`, durations uniformly sampled from fixed set: 1 day, 2 days, 3 days, 4 days, 5 days, 6 days, 7 days, 8 days, 9 days, 10 days, 1 week, 2 weeks, 3 weeks, 4 weeks, 7 days, 10 days, 14 days, 21 days, 25 days, 30 days, 1 month, 2 months, 3 months, 4 months, 6 months, 8 months, 4 weeks, 6 weeks, 8 weeks, 10 weeks, 1 year, 2 years, 3 years, 4 years, 12 months, 18 months, 24 months, 36 months.
`notable`	Uniformly sampled from a fixed set, extracted from Wikipedia. Omitted from brevity, full dataset available in the code repository.
`time_of_day`, `time_of_day_phase`	Action time uniformly sampled from all hours at :00, :15, :30, :45. Reference time sampled uniformly from all times of the day.

Table 8: BERTScore-based variability statistics for our dataset compared to prior datasets commonly used in studies of representational geometry and the linear representation hypothesis. Our dataset exhibits substantially higher variability across similarity pairs.

Dataset	Mean Similarity	Std
Our datasets		
cities_3way	0.8170	0.0217
date_3way_season	0.8306	0.0244
date_3way_temperature	0.8480	0.0240
date_3way	0.8218	0.0250
duration_3way	0.7975	0.0311
notable_3way	0.8099	0.0281
periodic_3way	0.8023	0.0302
time_of_day_3way_phase	0.8169	0.0265
time_of_day_3way	0.8174	0.0257
Heinzerling & Inui (2024) and El-Shangiti et al. (2025)		
P569 birthyear	0.8582	0.0444
P570 death year	0.8485	0.0517
P625.lat latitude	0.8289	0.0489
P625.long longitude	0.8598	0.0409
P1082 population	0.8544	0.0430
P2044 elevation	0.8410	0.0445
Engels et al. (2025)		
days of week	0.9627	0.0171
month of year	0.9617	0.0144

task have been obtained from Wikipedia[5] and have been rewritten via ChatGPT and checked manually to ensure consistence. For all datasets, each sentence contains exactly one temporal expression. We chose not to include more, using composite expressions, so as to obtain cleaner feature manifolds. The only exceptions are the `duration` and `time_of_day` datasets that contains two. This was necessary in order to formulate non-trivial questions that require reasoning across time spans. The `notable` task not only requires comparing different expressions but also involves factual recall of events from parametric memory. Variants `date_season`, `date_temperature`, and `time_of_day_phase` contain the same contextual sentences as the original tasks but a different termination that elicits a classification-based form of reasoning. Finally, we note that the number of examples per dataset varies. This is necessary to ensure that, after filtering for correctly answered instances, a sufficient number of activations remain for SMDS training. For consistency, we require a minimum of 500 correctly classified examples per model-task pair and cap the number of activations used in manifold search at this threshold. See Table 7 for a more extensive collection of examples.

B.1 Data Variability

Overall, as shown in Table 8, the dataset we created demonstrates higher variability compared to datasets curated by prior work in related areas. In particular, we compute several BERTScore-based (Zhang et al., 2019) variability metrics across all splits of our dataset and compare them against the dataset used in Heinzerling & Inui (2024) and El-Shangiti et al. (2025), as well as the dataset in Engels et al. (2025), works that established the study of representational geometry and the linear representation hypothesis. We measure variability by computing BERTScore F1 for every possible pair of texts in a dataset split, using the model's predicted-reference pairs formed via 5000 randomly sampled combinations. The mean similarity is the average of these pairwise F1 scores, while the std is the sample standard deviation of the same set of scores.

C Supervised Multi-Dimensional Scaling

In this section we provide further details on the dimensionality reduction method we use throughout the paper, as well as highlight its differences and similarities with other techniques which served as inspiration.

Description SMDS is based on the assumption that points $X \in \mathbb{R}^{n\times d}$ in the residual stream roughly lie on a feature manifold that can be parametrized with labels $y \in Y$ with $y_i \in [0, 1]$. Distances on this ideal feature manifold are assumed similar to Euclidean distances in the residual stream. Formally, given activations $X \in \mathbb{R}^{n\times d}$, two samples $x_i, x_j \in X$ and a linear projection $W \in \mathbb{R}^{m\times d}$ from the full space to the manifold subspace, we assume that $d(y_i, y_j) \approx \|W(x_i - x_j)\|$. To find W, we can minimize Eq. 1, reported here for readability:

$$\mathcal{L} = \sum_{i<j} \left(\|W(x_i - x_j)\|^2 - d(y_i, y_j)^2\right)^2.$$

The problem is solved as follows. First, ideal distances $d(y_i, y_j)$ between labels are computed and the squared distance matrix is defined as:

$$D_{ij} \coloneqq d(y_i, y_j)^2. \tag{4}$$

Then, classical MDS is performed. Double centering is applied:

$$H \coloneqq I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top, \quad B \coloneqq -\frac{1}{2}HDH \tag{5}$$

B is eigen-decomposed and a low-dimensional embedding Y is obtained:

$$B = V\Lambda V^\top \tag{6}$$

$$Y \coloneqq V_m\Lambda_m^{1/2} \in \mathbb{R}^{n\times m}, \tag{7}$$

with V_m the top m eigenvectors and Λ_m the corresponding eigenvalues. The embeddings Y represent the locations of data points in the parametrized approximation of the manifold such that $\|Y_i - Y_j\| \approx d(y_i, y_j)$. The following steps perform regression to find a mapping from datapoints X to this subspace. We center X and Y by subtracting their mean:

$$X_c = X - \overline{X}, \quad Y_c = Y - \overline{Y}. \tag{8}$$

Substituting in Eq. 1:

$$\|X_cW^\top WX_c^\top - Y_cY_c^\top\|^2. \tag{9}$$

At the optimum W we get that $X_cW^\top \approx Y_c$. Solving Eq. 1 is expensive, however we can approximate W by solving a proxy problem:

$$W = \arg\min_{\hat{W}} \|X_c\hat{W}^\top - Y_c\|. \tag{10}$$

this is the same formulation of a linear probe, but with Y being computed from the label using MDS. A solution is easily found:

$$W = Y_c^\top X_c \left(X_c^\top X_c\right)^{-1}. \tag{11}$$

[5] https://en.wikipedia.org/wiki/Timeline_of_the_20th_century

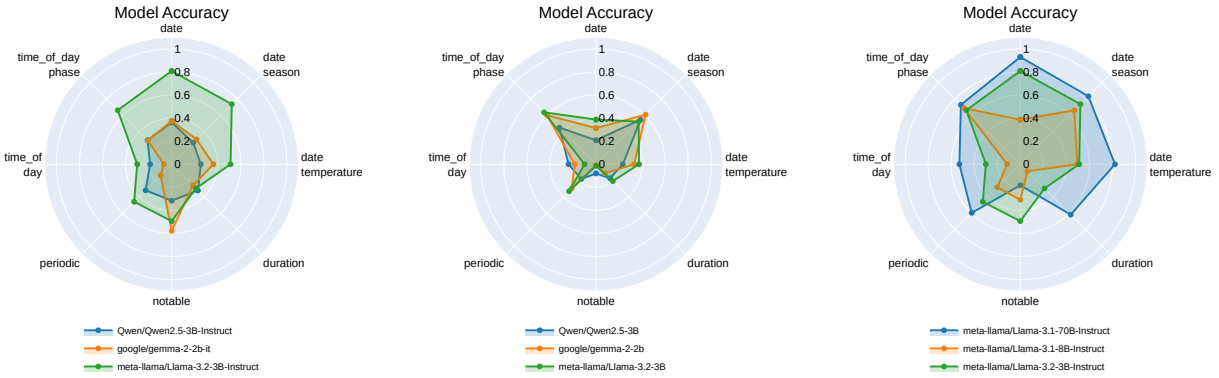

Figure 11: Accuracy on temporal tasks. Accuracy is low across the board, with only Llama models achieving above-chance accuracy. The 3B Llama variant is also shown outperforming the 8B one.

A regularization term can be added to Eq.11 to make the resulting projection more robust:

$$W = Y^\top X_c \left(X_c^\top X_c + \alpha I\right)^{-1}. \tag{12}$$

In all our experiments, we set $\alpha = 0.1$.

The embeddings Y represent the locations of data points in the parametrized approximation of the manifold (Figure 2). In principle, if such embeddings are already known, the preceding steps can be skipped entirely. Computing arbitrary $d(y_i, y_j)$ just gives more flexibility. By using SMDS to perform a search across candidate manifolds, the discovery problem can be reduced to one of model selection: one only needs to perform SMDS on several hypothesis metrics $d(y_i, y_j)$ or parametrized manifolds Y, and compare them using a quality metric like stress.

Comparison to other methods SMDS is an extension of MDS and uses it as part of the procedure. There is, however, a key difference in its use case: while classical MDS is unsupervised and only learns a lower-dimensional mapping that preserves Euclidean distances, SMDS first builds a distance matrix from labels and then uses it to learn the actual projection via regression. There are also differences in the stress metric we use (Eq. 3): classical normalized stress (Amorim et al., 2014) evaluates the error between distances in the original and lower-dimensional space; our formulation effectively does the same, but between the projected and the ideal subspace Y.

The first term in Eq. 1 can be reformulated:

$$\begin{aligned}\|W(x_i - x_j)\|^2 &= (W(x_i - x_j))^\top (W(x_i - x_j)) \\ &= (x_i - x_j)^\top W^\top W (x_i - x_j) \\ &= (x_i - x_j)^\top M (x_i - x_j),\end{aligned}$$

with $M \in \mathbb{R}^{n \times n}$ being a positive semi-definite matrix. This is the squared Mahalanobis distance, widely used in Distance Metric Learning. In fact, many other dimensionality reduction techniques can be described as Distance Metric Learning algorithms (Suárez-Díaz et al., 2020).

SMDS is closely related to probes, which have been extensively used in prior works (Belinkov, 2022; Li et al., 2022; Gurnee & Tegmark, 2023, *inter alia*). Some have successfully employed circular probes to recover feature manifolds (Engels et al., 2025) and study other cyclical patterns such as number encodings (Levy & Geva, 2025), but to the best of our knowledge no prior works have used MDS to build probes of arbitrary shape.

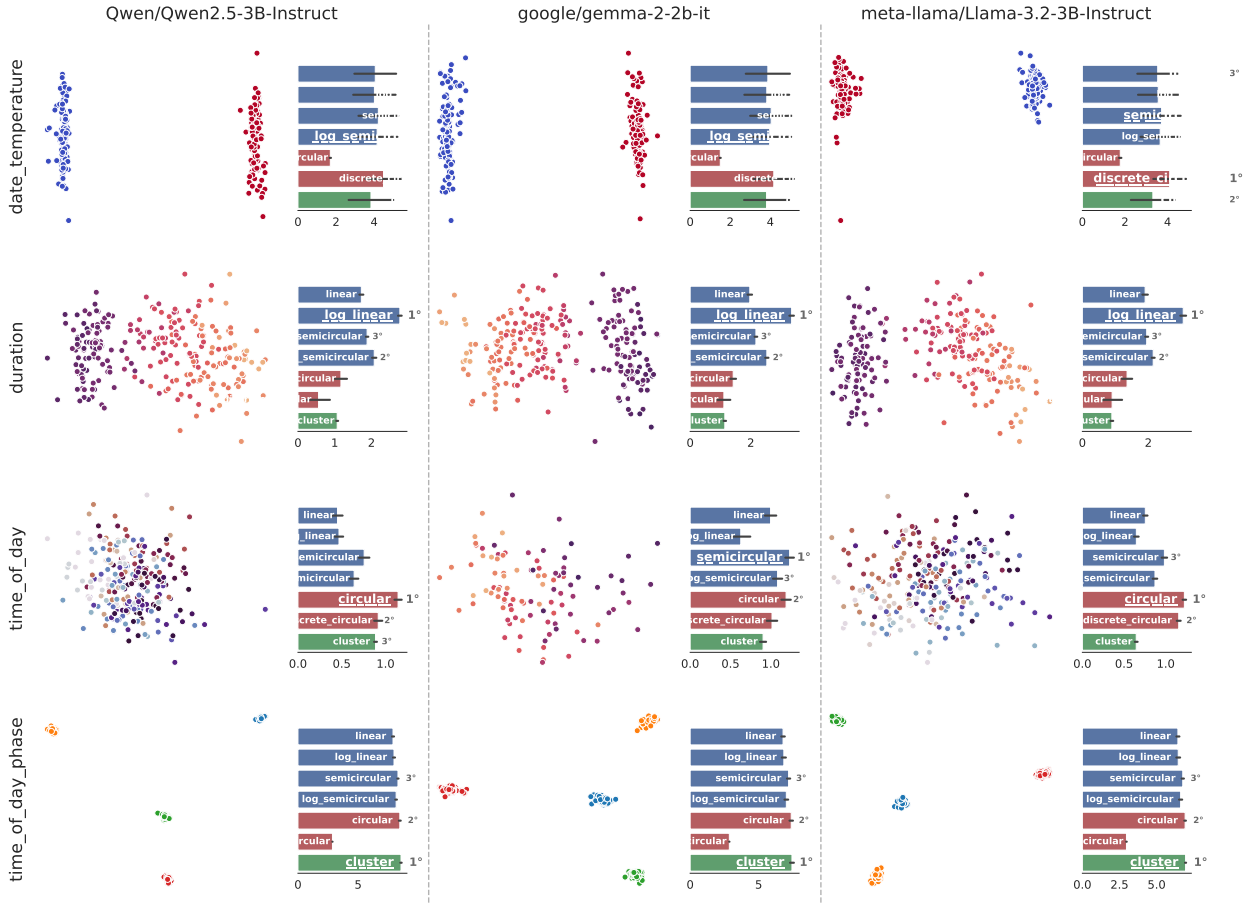

Figure 12: More manifolds for different tasks and models. Continued from Figure 3; error bars are shown in black. Manifold topology: linear; cyclical; categorical;

D Additional Experiments

D.1 Model Performance

We evaluate exact match accuracy across all tasks and models, finding that performance is generally very low, with only instruction-tuned models from the Llama family outperforming random chance. This is notable because, despite poor task performance, the models still produce well-defined feature manifolds. Among Llama models, we find the more recent Llama-3.2-3B-Instruct outperforms its 8B counterpart, while the 70B version displays stronger performance in almost all tasks (Figure 11).

While stress is a good indicator of performance, as discussed in Appendix 5.3, we observe no significant correlation for the three models of the main analysis (Spearman's $\rho = -0.034$). We hypothesize that while most LMs effectively structure knowledge internally, some struggle to leverage it during generation. This might explain their lower performance. Another possibility is that the specific wording of the prompt does not allow LMs to effectively recover information from context. In that case, chain-of-thought prompting (Wei et al., 2022) may improve performance.

D.2 Additional Observations on Manifold Discovery

We observe two instances where manifold discovery exhibits unexpected behaviours. On the `date_temperature` task (Figure 12), the clusters are correctly identified but the scoring yields unreliable values. This is expected when considering how distances are computed in the binary cluster scenario: two clusters can be modelled

correctly by any hypothesis manifold, as there is no order that can be enforced. This signals caution and suggests reverting to simpler probes to evaluate binary features.

On the `time_of_day` task, SMDS is unable to recover well-organised manifolds at the LP site despite a clear, preferential circular manifold being present at the TE site (Figure 15). The lack of transferability between the two sites can be explained by noting that `time_of_day` sentences contain two temporal expressions in the same format instead of one. The two representations may interfere destructively, preventing their recovery. When also considering findings from §D.3, it is also possible that, for this specific prompt, the LP site is not storing any semantically relevant information. Future works could start from tasks such as this to characterise how multiple feature manifolds combine, and in which token is this information encoded.

D.3 Additional Observations on Intervention

Figure 16 shows how the `time_of_day` is the least affected by intervention, even when perturbing the full latent space. We believe this is due to the specific formatting of time used: expressions such as `19:37` are tokenized as `19`, `:`, `37`, with the TE site corresponding to the minute part of the expression. For most examples, the hour is sufficient to determine the right answer, and since that information is left untouched, the model is able to continue with minimal disruption.

D.4 Identifying a Spatial Manifold

To show the flexibility of SMDS, we extend our analysis to a spatial reasoning task. In the same vein as Appendix B, we build sentences composed of three statements "`<name> lives in <city>.`" Then, we prepend a continuation "`The person who lives closest to <name> is`" to elicit reasoning. Names are sampled from the usual set, while cities are obtained from the World Cities Database[6]. We select only prominent cities as they are more likely to be present in the model's memory. For the US and Canada we instead select cities with > 10.000 inhabitants, since following the provided labels results in severe undersampling. We then uniformly sample cities based on location.

Each city is characterized by its latitude and longitude coordinates $c_i = (\text{lat}_i, \text{lon}_i)$. From these, we project cities on various shapes and compute the relative distance function. We investigate a flat plane, a sphere, a cylinder, and a complex geometry defined by the geodesic distance between cities. The flat manifold is computed simply as the Euclidean distance between the two coordinates, same as the `linear` metric used before. For the sphere manifold, we convert each coordinate into a 3D point on a sphere of radius r as follows:

$$\begin{aligned}
\phi_i &= \text{radians}(\text{lat}_i), \quad \lambda_i = \text{radians}(\text{lon}_i) \\
x_i &= r\cos(\phi_i)\cos(\lambda_i), \\
y_i &= r\cos(\phi_i)\sin(\lambda_i), \\
z_i &= r\sin(\phi_i)
\end{aligned}$$

Then the distance between two cities is the Euclidean chord length:

$$\|\delta_{ij}\| = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}$$

For the cylinder manifold, we map latitude to vertical height and longitude to angle around a cylinder of radius r. Each point is embedded as:

$$\begin{aligned}
h_i &= \text{radians}(\text{lat}_i) \cdot s, \quad \lambda_i = \text{radians}(\text{lon}_i) \\
x_i &= r\cos(\lambda_i), \\
y_i &= r\sin(\lambda_i), \\
z_i &= h_i
\end{aligned}$$

The chord distance is again computed as the Euclidean distance in 3D. For the geodesic manifold, we compute the great-circle distance between two cities—i.e., the shortest path along the surface of a sphere. We first

[6] https://simplemaps.com/data/world-cities

Table 9: Stress values for control tasks. Absolute difference with the base task is shown in red. SMDS consistently produces low scores if no structure is present.

Dataset	Llama-3.2-3B-Instruct Best shape	Llama-3.2-3B-Instruct $-\log S$	Qwen2.5-3B-Instruct Best shape	Qwen2.5-3B-Instruct $-\log S$	gemma-2-2b-it Best shape	gemma-2-2b-it $-\log S$
date	circular	$0.995_{-0.932}$	circular	$0.855_{-2.075}$	circular	$0.806_{-1.563}$
date_season	cluster	$0.998_{-4.509}$	cluster	$0.956_{-7.205}$	cluster	$0.858_{-6.308}$
date_temperature	cluster	$0.376_{-5.669}$	cluster	$0.332_{-7.010}$	cluster	$0.211_{-6.111}$
duration	log_linear	$0.513_{-2.490}$	log_linear	$0.519_{-2.083}$	log_linear	$0.445_{-2.059}$
notable	semicircular	$0.722_{-1.248}$	semicircular	$0.740_{-0.801}$	semicircular	$0.482_{-1.168}$
periodic	log_linear	$0.523_{-2.888}$	log_linear	$0.502_{-3.150}$	log_linear	$0.235_{-3.470}$
time_of_day	circular	$0.987_{-0.223}$	circular	$0.978_{-0.066}$	semicircular	$0.683_{-0.468}$
time_of_day_phase	cluster	$0.961_{-4.710}$	cluster	$1.000_{-6.869}$	cluster	$0.902_{-6.487}$

convert latitude and longitude to radians and compute the differences:

$$\phi_i = \text{radians}(\text{lat}_i), \quad \lambda_i = \text{radians}(\text{lon}_i)$$
$$\Delta\phi = \phi_i - \phi_j, \quad \Delta\lambda = \lambda_i - \lambda_j$$

We then use the Haversine formula:

$$\begin{aligned} a =& \sin^2\left(\frac{\Delta\phi}{2}\right) \\ &+ \cos(\phi_i)\cos(\phi_j)\sin^2\left(\frac{\Delta\lambda}{2}\right) \\ \|\delta_{ij}\| =& r \cdot 2\arcsin\left(\sqrt{a}\right) \end{aligned}$$

This corresponds to the true surface distance between two points on the Earth, assuming a perfect sphere. In addition to the usual LP and A sites, we analyze two more locations: the correct city (CC), corresponding to the final token of the city where the correct person lives, and the reference city (RC), referring to the final token of the city where the person in the question lives. Both cities are drawn from the context statements.

Table 5 shows that the manifold achieving the closest fit is a spherical one across all models. In Figure 17, we visualize the projection recovered by SMDS and find clear clusters around the shapes of continents. Their relative position is consistent with their real-life location, but projecting a spherical manifold onto a plane inevitably distorts their real position.

D.5 Feature Manifolds are not Artefacts of SMDS

In this section we validate the robustness of SMDS and confirm that the feature manifolds recovered are indeed consistent and not an artefact of overfitting a projection. Primary evidence is provided in the main analysis of §5: the error bars produced by cross-validation are narrow for almost all datasets, confirming that activations for a given feature do have a preferential manifold.

The second piece of evidence is obtained by designing control tasks following Hewitt & Liang (2019). We build control variants for all tasks by shuffling the labels. This should make it impossible for SMDS to identify a structure and we should observe a significant increase in stress. For each model-task pair, we evaluate the best manifold identified in §5. As in the main experiment, we perform a 5-fold cross-validation on the dataset. Table 9 shows the results: absence of structure causes a sharp increase in stress (and corresponding drop in $-\log S$). This is evidence that SMDS does not force a structure when no underlying manifold exists.

D.6 Exploring the Impact of Instruction Tuning

Since we exclusively use instruction-tuned models in our main experiments, we are interested in whether instruction tuning impacts the feature manifolds and accuracies of our models. Instruction tuning is a post-training method that is widely believed to enhance models' generalization and task-solving capabilities (Wei et al., 2021; Chung et al., 2024), however not all of the potential changes induced in a base model by

instruction tuning have been explored. Various prior works suggest that instruction tuning mainly impacts stylistic output tokens rather than changing the model's parametric knowledge (Zhou et al., 2023; Ghosh et al., 2024; Lin et al., 2023), and that it additionally causes models to rotate the basis of their representation space to adapt to user-oriented tasks (Wu et al., 2024). However, these works deal primarily with token probability distributions and do not explore feature manifolds.

To explore the impact of instruction tuning in our experimental setup, we conduct our main experiments on the base versions of three of our models: Llama-3.2-3B, Qwen2.5-3B, and Gemma-2-2B. We report the stress values in Table 3 and the feature manifolds in Figure 14.

Overall, across our three models, we find that instruction tuning did not substantially alter the optimal structure. For all tasks except `notable`, `periodic`, and `duration`, the structure that was optimal for the instruction-tuned model tended to remain in the top-3 optimal structures for the base model as well. We note that the three outlier tasks have monotonic topologies, while the rest have cyclical or cluster-like structures.

For some tasks, we observed that the feature manifolds for base models tended to be more scattered than those of the instruction-tuned models. For example, the `date` manifolds for all instruction-tuned models (Figure 13) have tighter, well-formed ring structures than the manifolds of the base models (Figure 14). However, this was not the case for all tasks: for example, the `periodic` manifolds for both base and instruction-tuned models showed a clear separation of clusters that remained consistent within a particular model architecture.

Surprisingly, we found that the accuracies varied unpredictably between the base and instruction-tuned models. For Llama, all base accuracies were much lower than the instruction-tuned accuracies, while with the Qwen and Gemma models, base models sometimes markedly outperformed the instruction-tuned models. This could possibly be due to differences in the instruction-tuning methods of these models.

Overall, our results suggest that the impact of instruction-tuning on feature manifolds will depend on the task, model architecture, as well as on the specifics of the instruction-tuning process. A detailed exploration of this is a promising avenue for future work.

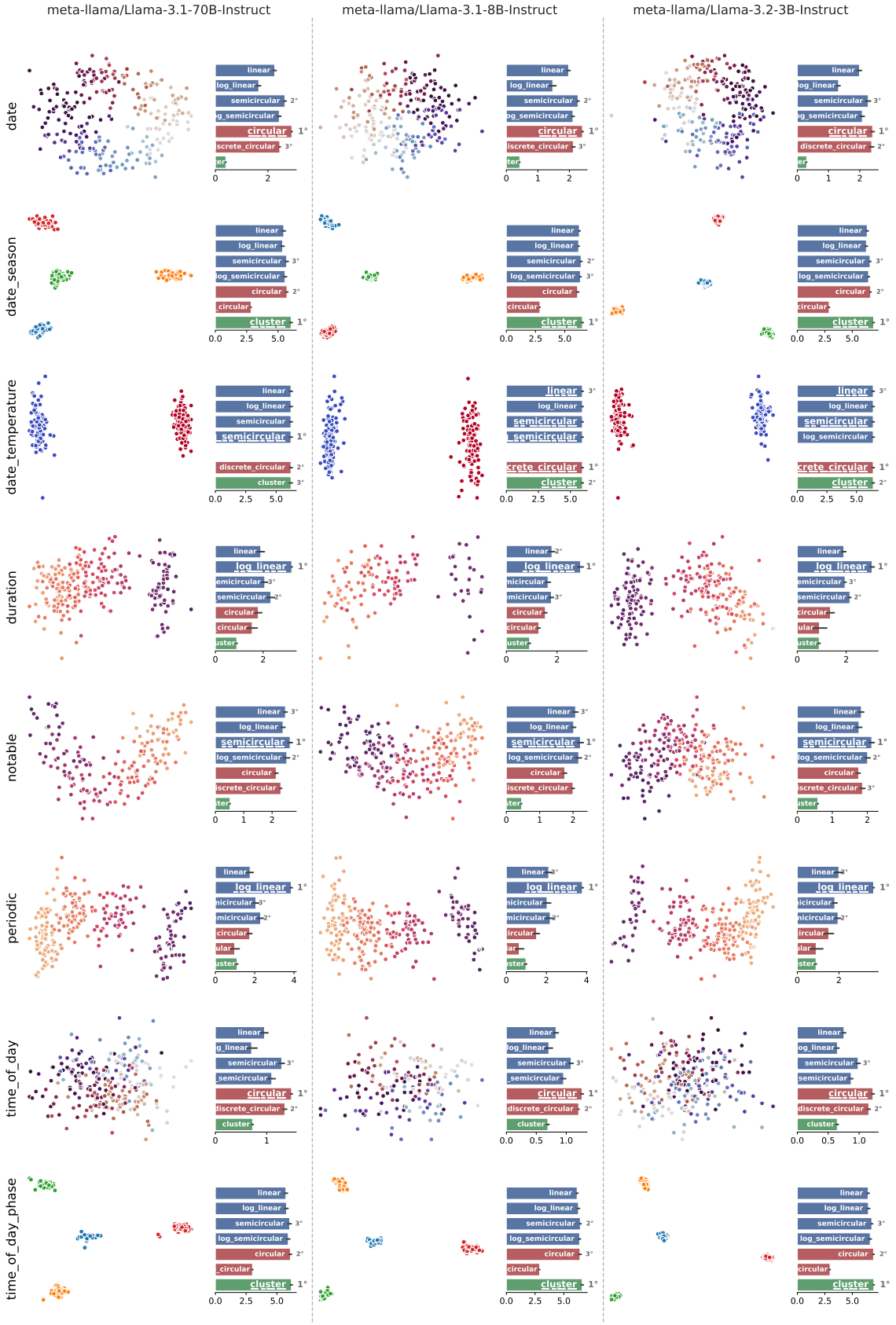

Figure 13: Feature manifolds for models at different sizes. There is a preferential manifold also across scales. Continued from Figure 3; error bars are shown in black. Manifold topology: linear; cyclical; categorical;

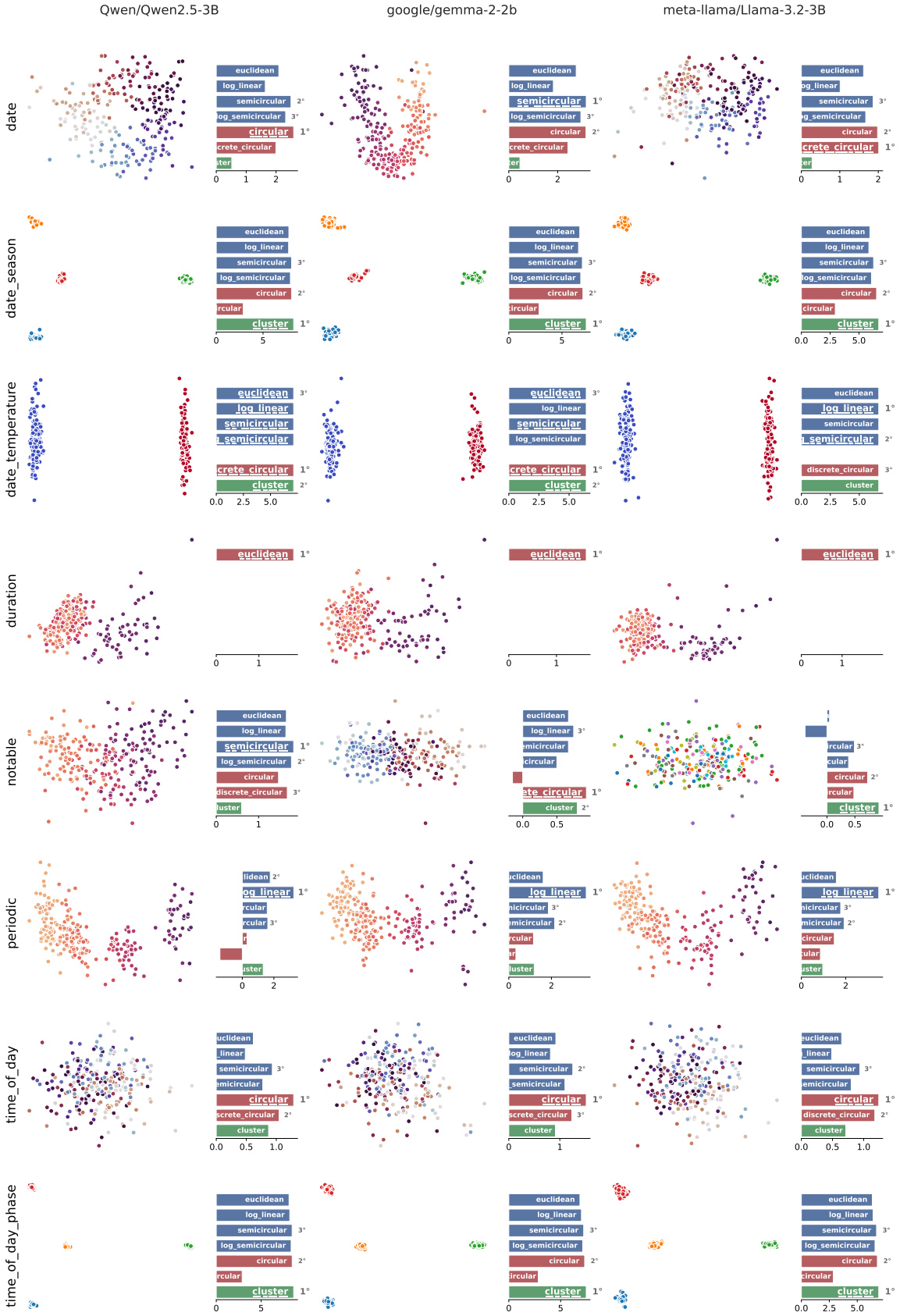

Figure 14: Feature manifolds for base models. Geometries are consistent with the instruction-tuned counterparts in most cases. Manifold topology: linear; cyclical; categorical;

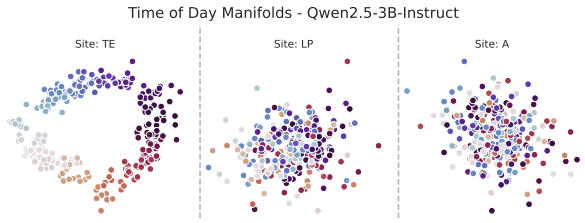

Figure 15: Circular manifolds on the `time_of_day` task. SMDS cannot find any structure on LP and A despite one being present at the TE site.

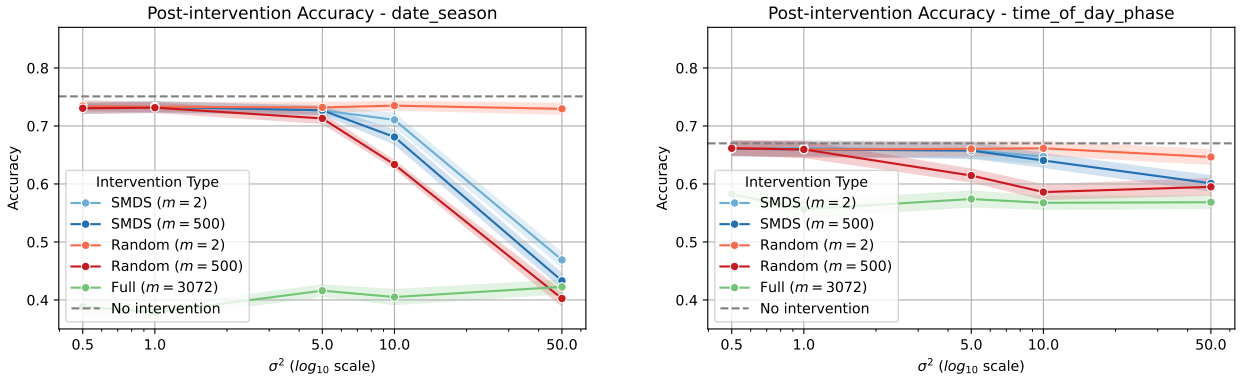

Figure 16: Additional accuracy plots from the intervention experiment. Error bars represent standard error. The `time_of_day` task is the least affected by all forms of intervention.

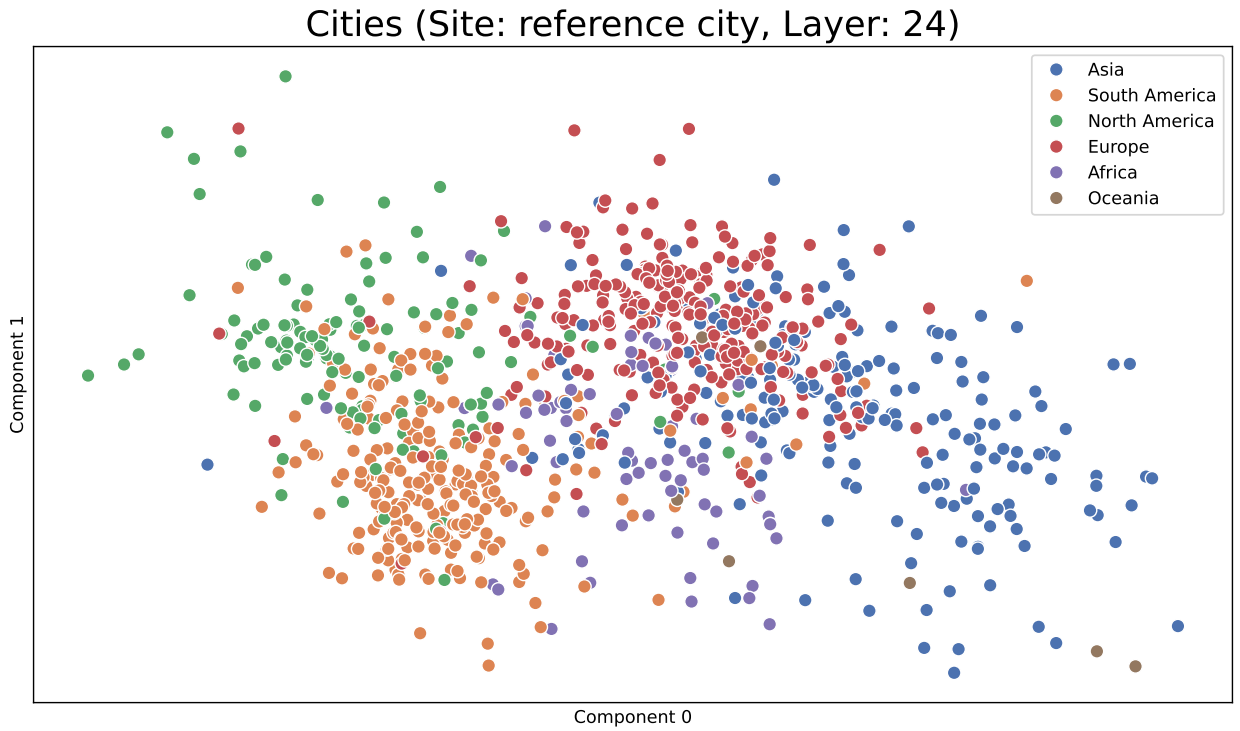

Figure 17: SMDS of gemma-2-2b-it on the `cities` task. The recovered projection shows the relative position of continents. For the sake of clarity, the flat manifold is shown instead of the best-scoring spherical one.